Univerzita Karlova v Praze

Filozofická fakulta

ÚSTAV ANGLICKÉHO JAZYKA A DIDAKTIKY

DIPLOMOVÁ PRÁCE

Bc. Simona Zvěřinová

N-gramy v mluveném projevu českých a rodilých mluvčích angličtiny

N-grams in the speech of Czech and native speakers of English

Praha, 2016

Vedoucí práce: PhDr. Tomáš Gráf, Ph.D.

Tímto bych chtěla poděkovat PhDr. Tomáši Gráfovi, Ph.D za bezmeznou trpělivost, cenné rady a připomínky a stejně tak i za nesčetná slova podpory.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Souhlasím se zapůjčením bakalářské práce ke studijním účelům.
I have no objections to the MA thesis being borrowed and used for study purposes.

V Praze dne 16. 8. 2016 _____

Simona Zvěřinová

Abstract

The diploma thesis is concerned with the analysis of recurrent word-combinations in the speech of advanced Czech speakers of English and native speakers of English. The data used for the analysis is extracted from two corpora, learner corpus LINDSEI and native speaker corpus LOCNEC. The aim of the thesis is to compare the two groups of speakers, determine differences in their use of recurrent word-combinations and compare the findings to previous studies involving speakers of different languages. The quantitative analysis is performed on a sample of 50 speakers from each corpus and the frequency data is used to compare the two groups as to the number of types of word-combinations they use and how frequently they do so. The qualitative analysis is performed on a sample of 15 speakers from each corpus to determine functional differences. Four categories of word-combinations are determined in the analysis. In the conclusion, the quantitative and qualitative findings are compared to previous research involving speakers of different languages.

Keywords: spoken language, learner language, n-grams, n-gram analysis, recurrent word-combinations, lexical bundles, learner corpus

Abstrakt

Diplomová práce se zabývá analýzou opakovaných slovních spojení v projevu pokročilých českých mluvčích angličtiny a rodilých mluvčích angličtiny. Data využitá v analýze jsou čerpána ze dvou korpusů, žákovského korpusu LINDSEI a korpusu rodilých mluvčích LOCNEC. Cílem práce je porovnat dvě skupiny mluvčích, odhalit rozdíly mezi jejich užíváním opakovaných slovních spojení a porovnat výsledky s předešlými pracemi zahrnujícími výzkum mluvčích jiných jazyků. Kvantitativní analýza je provedena na vzorku 50 mluvčích z každého korpusu a frekvenční data jsou užita k porovnání mluvčích na základě toho, kolik typů slovních spojení užívají a jak často. Kvalitativní analýza je provedena na menším vzorku 15 mluvčích z každého korpusu a určuje funkční rozdíly. Během analýzy jsou určeny čtyři kategorie slovních spojení. V závěru jsou kvantitativní i kvalitativní výsledky porovnány s předešlým výzkumem mluvčích jiných jazyků.

Klíčová slova: mluvený jazyk, žákovský jazyk, n-gramy, n-gramová analýza, opakovaná slovní spojení, lexikální svazky, žákovský korpus

Contents

List of abbreviations

LGSWE - Longman Grammar of Spoken and Written English

L1 – first language

L2 – second/foreign language

LINDSEI_CZ, LS – the Czech subcorpus of the Louvain International Database of Spoken

LOCNEC, LC – The Louvain Corpus of Native English Conversation

LSS – LINDSEI_CZ Sample

LCS – LOCNEC Sample

T1T2 – Task 1 and Task 2

CZ followed by number (e.g. CZ001) – speaker identification in LINDSEI_CZ

EN followed by number (e.g.EN001) – speaker identification in LOCNEC

FREQ, FQ – frequency

RNK – rank

RNG – range

N. – number of occurrences

List of tables

List of figures

1 INTRODUCTION

Interest in research on recurrent word-combinations and learner language started increasing only since the end of the last century, when the importance of phraseology in foreign language learning was properly noted back in the 1990s. The number of studies where phraseology is studied in the production of foreign speakers of English has multiplied in the recent years and the interest does not seem to be waning just yet. One of the reasons is that it has been made increasingly easier for researchers to perform extensive studies on large samples of data since the 1980s, when large collections of learner-language data started being collected, and since the creation of complex computerised learner-language corpora.

The importance of phraseological studies of learner language for the areas of language proficiency and language fluency has also proven to be immeasurable. More and more comparative studies of native language production and foreign language production have been conducted in fairly large numbers, many of them focused on or involving the investigation of recurrent word-combinations. Most studies first focused on written language because the data is easier to acquire and parse for the purpose of corpus tagging. Nevertheless, there have been comparative studies involving the spoken language production of speakers of many languages, for example German (Götz, 2013), French (De Cock, 1998 and 2004), Swedish (Aijmer, 2004) and Norwegian (Larsson Ass, 2011). The studies often reflect on previous research, making observations about features common to learners from different language backgrounds possible.

The main aim of the present thesis is to compare Czech and native speakers of English in terms of their usage of recurrent word-combinations. The speech of advanced learners appears to be similar to native speaker language production in many aspects and yet studies have shown considerable differences in their usage of recurrent word-combinations, be it underuse or overuse from a quantitative perspective or the use of certain types of word-combinations for purposes different from those of the native speakers. As the study is the first comparative study of recurrent word-combinations that involves Czech speakers, it not only aims to investigate differences between Czech speakers and native speakers in terms of frequency and function, the goal is also to reflect the findings back on research done on speakers of different L1 backgrounds.

The thesis consists of 5 chapters. Chapter 2 provides theoretical background to the study. It shortly introduces linguistic approaches to spoken English, the aspects of real-time production of language as observed by Geoffrey Leech (2000) and a short glimpse into the

study of learner language and fluency in the context of recurrent word-combination research. The chapter further discusses recurrent word-combinations within the wider framework of phraseology and also presents an overview of relevant research on recurrent word-combinations. Chapter 3 describes the spoken learner corpora employed in the study – the Czech subcorpus of the Louvain International Database of Spoken English Interlanguage and the Louvain Corpus of Native English Conversation. It also specifies the method of extraction and modification of the data used in the analysis and then shortly outlines the analysis conducted in the practical part of the thesis.

Chapter 4 contains the quantitative and qualitative analysis of recurrent word-combination of two samples of different size. The findings of the largely qualitative part of the analysis are summarised and discussed at the end the Chapter. The final chapter, Chapter 5, discusses the overall findings in the context of previous research, reflects on the shortcomings and limitations of the study and suggests possibilities for further research.

2 THEORETICAL BACKGROUND

2.1 Spoken language

Spoken language has often been described using traditional models of grammar applicable primarily to written English, often excluding features that are unique to spoken language. (Carter & McCarthy 1995: 141) However, with increasing availability of natural spoken data, attempts at remedying the shortcomings of this approach to the exploration of spoken English have been made.

Corpus-oriented studies are of descriptive nature, very much focused on language use rather than on the creation rules and categorisations based on abstracted language. Some theories, however, may make use of the empirical findings of such studies in that they hold the view that "grammar as a mental system is mirrored closely in the way language is used, and some (e.g., probabilistic grammars) can scarcely be tested or formulated without resort to a corpus." (Leech, 2000: 686) Leech also claims that it can be argued that since the aim of learners of foreign languages is to productively and receptively communicate in their chosen language, the performance grammar observed in natural spoken data is an invaluable source for language learners and teachers.

Leech speaks of two approaches to spoken language (mainly grammar) that have crystallised with increasing access to spoken corpora (Leech, 2000: 687):

Approach A: Emphasizes the differentness of spoken grammar from previously articulated grammatical models.

Approach B: Asserts the underlying sameness of spoken and written grammar along with notable differences in frequency.

According to Leech, one of the most prominent figures of Approach A, Michael McCarthy, while anticipating common ground between spoken and written grammar, "goes so far as to argue that there should be no prior assumption that the grammar of speech and the grammar of writing share the same framework (...) and has argued for a different model of grammar for speech (...) [and] for a close integration of spoken grammar and discourse analysis." (Leech 2000: 688)

The prominent figures of Approach B, Biber et al. (1999), in their *The Longman Grammar of Spoken and Written English*, on the other hand, describe spoken language in terms of lexico-grammatical features, using terminology and framework based on Quirk et

al.'s *Comprehensive Grammar of the English Language (1985)*, assuming a common model for written and spoken English alike. However, Biber et al.'s grammar also presents a more performance-based view of spoken language, showing "very marked differences of frequency in the way grammar is used in speech and in writing." (Leech 2000: 690) They also present a functional interpretation of their findings largely dependent on the spoken nature of the data. The function of a feature in LGSWE takes the following three forms (Biber et al. 2000: 41):

(1) the work that a feature performs in discourse
(2) the processing constraints that it reflects
(3) the situational and social distinctions that it conventionally indexes

Biber et al. then consider contextual circumstances of spoken discourse and also the production and comprehension circumstances of certain features of language, reflecting on constraints placed on speakers in real-time production.

Aside from Biber et al.'s admittedly sweeping consideration of language production, in the above publication, there has been more research done into the area which considers performance and production. Biber himself, along with Susan Conrad (2005), considers performance factors and processing pressures in his study *The Frequency and Use of Lexical Bundles in Conversation and Academic Prose*. Previous research into formulaic language done by Wray and Perkins (2000) also shows devices which help balance processing difficulties involved in real-time language production while also exploring formulaic features of spoken language as a tool for social-interaction.

A mention must be made of approaches which consider features unique to spoken language such as back-channelling, turn-taking, and various other sound-based aspects (eg. intonation) in their analyses, usually with the help of prosodic transcription when a corpus of any kind is involved. The fields taking such an approach – for example conversational analysis or interactional sociolinguistics – are approaches involved precisely in the field of discourse analysis, whose importance McCarthy emphasises in connection to the study of spoken English. Other studies which tend to consider contextual features of spoken language, its sound-based aspects and which take a functional view of spoken language features are studies involving research into foreign language acquisition, learner language and language proficiency (see more in chapter 2.1.2).

2.1.1 Conversational grammar

As the study works with two corpora of natural spoken data of spoken English, it is imperative to provide a kind of summary of some prominent aspects of language produced in real-time. Leech discusses in his work the findings collected in Biber et al.'s LGSWE and gives a list of circumstances which determine the nature of conversational grammar, applicable to the nature of conversation itself. (Leech, 2000: 701)

(1) Conversational grammar reflects a shared context
(2) Conversational grammar avoids elaboration or specification of reference
(3) Conversational grammar is interactive grammar
(4) Conversational grammar highlights affective content: Personal feelings and attitudes
(5) Conversational grammar has a restricted and repetitive lexicogrammatical repertoire
(6) Conversational grammar is adapted to the needs of real-time processing

To show how all these points interact in conversation, Leech provides an explanation (with the aid of Figure 1): *Interactive* dialogue enables grammatical shortcuts on the basis of ongoing *shared context*; *shared context* causes the tendency to rely on implicit reference which requires no elaboration; *low specification* caused by the lack of need to elaborate and specify means that the speaker can rely on a repetitive repertoire of words and phrases; *real-time processing* or rather pressures connected with it encourage reliance on a *limited repertoire* of items readily retrievable from memory; *interactiveness* clearly associates with *affectivity*, each involving personal and experiential aspects of communication. (Leech, 2000: 701)

The principles of shared context and real-time processing are, according to Leech, the two key situational factors which explain the functional nature of conversation, and also the two that are most independent of language (cf. low specification and restricted repertoire).

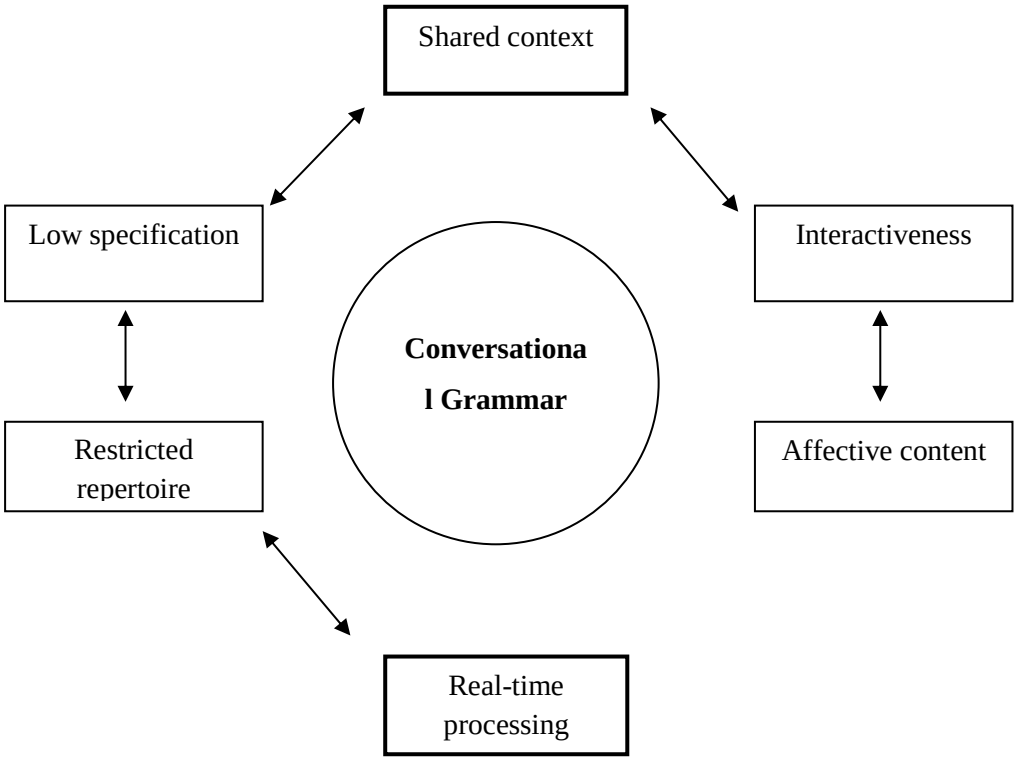

Figure 1: Interrelated functions associated with conversational grammar (Leech, 200: 701)

This observation partly overlaps with one which appears in a study on the function of formulaic language written by Wray and Perkins (2002) (more on this study in chapter 2.2.2.1). This study observes that formulaic language, which is in the centre of it, may and does function as a tool during real-time conversation; as a device of *social interaction* and as *compensatory devices for memory limitations*. (Wray and Perkins, 2000: 14, 15) Word-combinations which function as devices of social interaction, such as politeness markers, are clearly linked to Leech's functions of interactiveness and affective content. Compensatory devices such as *processing shortcuts* or *time-buyers* can also be observed in Leech's function of real-time processing. Word-combinations, included as such in Leech's function of restricted and repetitive repertoire, can be seen in Wray and Perkins' study to be involved in many of Leech's other functions of conversational features.

2.1.2 Learner language and proficiency in the context of recurrent word-combinations

The interest in learner language and the desire to pick it apart and study it is only a few decades old. In the period spanning 1960s and 1970s, the common understanding of language was reached; there were three language systems of interest to the study of the

process of second language acquisition: native language, target language and **learner language**. This was in opposition to the widely held behaviourist belief that the second language of a learner took a certain form influenced by their first language; owing to this belief, the contrastive analysis of native language and target language was expected to uncover all difficulties learners might encounter in their acquisition of the second language. S. P. Corder and L. Selinker coined the terms *transitional competence* and *interlanguage*, respectively, and introduced the idea of "an autonomous linguistic system in its own right that evolved according to innate and probably universal processes," believing that "cognitive processes other than transfer might be at work in shaping this third linguistic system." (Tarone, 2014: 8)

Natural learner language data then presents an opportunity to glimpse into this third system; the present study acknowledges the idea of interlanguage and approaches it on its own while also viewing it in contrast with the native form of the target language. The two methodological maxims concerning the study of interlanguage semantics are as follows (Selinker, 2014: 234):

(1) Any interlanguage data should be considered idiosyncratic until shown to be otherwise;
(2) No matter how advanced the interlanguage speaker, there will exist both similarities and differences between interlanguage form/meaning combinations and those of the target language.

The second point concerning the mappings between from and meaning combinations have already proven to be relevant in research into formulaic learner language, where the discursive functions of certain formulaic sequences have shown to be different for foreign learners and for native speakers (see more in 2.2.2.1).

Another, arguably one of the most prominent concerns of SLA research, is research into second language **proficiency**. The idea of proficiency brings along the rather difficult task of identifying what it means to know a language. In practice, the meaning and value of proficiency is varied; SLA researchers view proficiency as evidence of L2 development, L2 teachers as the state of development which affects teaching practice, for L2 testers it is a gauge of achievement, for the public a guarantee of an individual's future achievements and for the learner it is a combination of all of the above – it provides the learner with key information for further learning and influences motivation. (Gráf, 2015: 17)

The following few paragraphs introduce, in short, a certain view of proficiency, in which the concept of fluency plays a significant role. Their primary function is to put the concept of fluency into context, as it plays a significant part in the background of this study and the following research into recurrent word-combinations may show some, if fairly small, findings relevant to fluency research.

One of the questions at the centre of SLA research into proficiency is whether it is realistically possible to measure proficiency using linguistic means. There are three dimensions that have been recognised as key components of proficiency and performance: **complexity**, **accuracy** and **fluency**. (Gráf, 2015: 18) The concepts have been separately researched since the 1960s and the concepts of accuracy and fluency were later combined in the context of communicative language classroom by Brumfit (1984) who believed that learners could not produce L2 effectively while focusing both on fluency and accuracy and therefore have to sacrifice one for the benefit of the other. Brumfit's theory brought the dichotomy of fluency-oriented and accuracy-oriented activities into language classrooms. "However sound this advice appears to be for classroom practice it failed to explain the inter-relation between the two concepts, neither did it attempt to initiate a discussion about the definition of accuracy and fluency." (Gráf, 2015: 18)

In the second half of 1990s, Skehan (1996; 1998) proposed a model of proficiency which later became an influential force in the field of SLA. Owing to its consideration of all three key components, the model became known as the CAF model. In the context of this model, the three features accompany cognitive and psycholinguistic processes; "these depend on automaticity, parallel and controlled processing, proceduralisation, conscious awareness, use of attentional resources, type and speed of processing, difficulty and relative novelty of tasks, declarative and procedural knowledge, memory and retrieval." Similarly to Brumfit's claim where fluency and accuracy suffer as the speaker focuses their mental faculties more on one or the other, Skehan (1998) claims that "speakers' information processing capacity is limited and as a result speakers have to choose to which of the dimensions they pay more attention." (Gráf, 2015: 20)

2.1.2.1 Brief introduction into the study of fluency

While research into formulaic language, specifically into recurrent word-combinations, may have bearing on other areas of proficiency, it plays a significant part in research into language fluency. It has been pointed out that fluency has often been only very vaguely defined. Wood (2010), for example, speaks of this matter as follows.

> In general parlance, fluency is often used as a synonym for effective spoken use of a language. It is frequently used to mean "native-like," having a high overall degree of proficiency, or having a "good command" of a language. In the language teaching profession, fluency is generally more tightly defined. We tend to use the word to mean a naturalness of flow of speech, or speed of oral performance. (Wood, 2010: 9)

Research into fluency has played part in various fields, from psycholinguistic research to empirical research into production and temporal aspects of speech. Its relevance has also been proven in research in the field of SLA and language teaching. The large body of research into language competence and proficiency has shown that fluency is influenced by a great number of competencies. (Wood, 2010: 10)

This research on fluency falls into four large categories: temporal variables of speech; the nature and functions of formulaic language; automatization and mental processing in language production; social and cultural aspects of fluency. (Wood, 2010: 9) Wood (2010), looking into previous research into these four large categories summarises the aspects most relevant in the research on fluency. The categories most relevant to the present study are the temporal variables and formulaic language. Formulaic language is explored within the framework of previous contrastive interlanguage research in chapter 2.2.2.1 and therefore the following paragraphs focus solely on briefly introducing the role of temporal variables in studies of fluency.

Concerning the temporal variables of speech, researchers seem to have a high degree of agreement as to which of them are most relevant. These are:

(1) Rate of speech
(2) Repair phenomena
(3) Pause phenomena
(4) Length of fluent runs

The *rate of speech* variable has shown to be relevant to fluency. It is usually measured as syllables uttered per minute or second. Speech rate seems to increase along with other measures over time and with the increased rate a speaker tends to be perceived as more fluent also. While the rate of speech has shown to be a sound indicator of fluency, "it seems that

speed gives us little information about the workings of fluency unless it is viewed in interaction with certain other variables." (Wood, 2010: 20)

Repair phenomena such as self-corrections and repetitions have shown, according to Wood (2010), mixed results in the research into fluency. Perceived fluency has not been shown in research to correlate in any conclusive terms to repair phenomena, as some people perceived as fluent speakers have shown to sometimes use more of them than others who were perceived as less fluent due to other features of their speech. "It appears that while repair phenomena may have something to tell us in qualitative terms about how fluency develops or occurs, repairs are only weakly linked in the literature with overall development of fluency." (Wood, 2010: 22)

Pause phenomena are "the most complex and one of the most informative elements of fluency." (Wood, 2010: 23) Frequency and location of pauses are the two aspects of fluency that have been most studied. While length and frequency of pauses has an impact on perceived fluency, it is the location of pauses which informs researchers more on the nature of the relationship between fluency and psycholinguistic mechanisms of production. The location of pauses in speech is an important indicator of fluency. The clustering of pauses signals reduced fluency and the syntactic location is just as salient. "Highly fluent second-language speakers and native speakers tend to pause at sentence and clause junctures, or between non-integral components of clauses and clauses themselves. Pausing at other points within sentences and clauses gives the impression of disfluency." (Wood, 2010: 27)

Length of fluent runs is the most important indicator of fluency discovered to date. It does not only influence perceived fluency, it also provides a key with which to facilitate the development of fluency through instruction. This temporal variable is closely connected to recurrent word-combinations; an increasing blend of automatized chunks of formulaic strings and frameworks of speech, together with newly assembled strings of words seems to enable speakers to produce the longer runs between pauses. (Wood, 2010: 29) The ability to work with automatic chunks of language, encoding and producing them, and the generating of new words and constructions puts a lot of pressure on a speaker. The process of this complicated planning seems to happen on two levels (Wood, 2010: 30):

(1) topic and overall syntax structure are planned in advance chunks, ideally identical with the clause/statement breakdown of the passage. In practice, however, the subjects *are* forced to break these units down still further . . .

(2) Planning at the level of lexical selection would appear to be on more of an ad-hoc basis. The self-corrections at this level would indicate the late stage at which this planning takes place.

2.2 Recurrent word-combinations

Two areas actively involved in research into recurrent word-combinations relevant to the present study are the field of linguistic phraseology and the field of SLA, focusing on learner language, often through contrastive interlanguage analysis. This chapter shortly presents the wide scope of views of recurrent word-combinations, their place amongst linguistic theories, some terminology concerning this phenomenon and the previous research more or less closely connected to the present study done in both fields. Studies from both fields often overlap and therefore are not strictly divided between the two separate chapters.

2.2.1 Phraseological and frequency-based approach to word-combinations

> Although the study of multi-word units has a long history, with Bally distinguishing between the fully fixed 'unités phraséologiques' and the looser 'séries phraséologiques' as early as 1909, phraseology has only recently begun to establish itself as a field in its own right. This process is being hindered by two main factors however: the highly variable and wide-ranging scope of the field on the one hand and on the other, the vast and confusing terminology associated with it. (Granger and Paquot, 2008: 27)

Phraseology has been previously defined as "the study of the structure, meaning and use of word combinations." (Cowie 1994: 3168) These word-combinations, however, are specifically studied as phraseological units. The criteria for what should be considered a phraseological word-combination vary from approach to approach; in general, it is possible to observe two major traditions concerning the issue of distinguishing which word-combinations can be considered phraseological and which cannot.

The traditional, **phraseological** view, finds its roots among the scholars of the former Soviet Union and other countries in Eastern Europe. In this tradition, the wide field of

phraseology becomes restricted to a specific subset of varied multi-word units defined in linguistics and becomes a continuum of word combinations from the opaquest, semantically opaque and elephant fixed, to the most transparent and variable ones. One of the key concerns of the phraseological approach is to find linguistic criteria according to which it would be possible to distinguish phraseological units one from another, with the most idiomatic units at the field's core and the least idiomatic at its periphery or, indeed, outside of it. (Granger and Paquot, 2008: 28)

The **frequency-based** approach is more recent and takes a bottom-up, inductive view of language instead of a top-down view based on pre-established criteria. At the forefront of this approach is John Sinclair's idiom principle; a view of language which suggests that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appeal to be analyzable into segments." (1991: 100) Phraseology takes a central place in Sinclair's model of language and many units falling to the periphery in the view of the traditional phraseological approach also become central, owing to the fact that they have shown to be more prevalent in language than fixed combinations such as idioms or proverbs. (Granger and Paquot, 2008: 29)

More specific and fairly crucial differences between the two approaches also become evident in their view of word-combinations in the context of semantics, morphology, syntax and discourse. Very succinctly and generally said, the traditional phraseological approach to semantics tends to exclude from consideration any units that are completely compositional and their meaning can therefore be clearly derived from the meanings of all its parts. This goes hand in hand with the fact that the traditional phraseological approach also observes a difference between the free combinations only governed by semantic co-occurrence and multi-word combinations whose meaning cannot be derived using semantics and therefore does not see the former as part of the field of phraseology. (Granger and Paquot, 2008: 30, 31)

The frequency-based approach takes a different view of meaning, claiming that the meaning of a word extends beyond the limits of the actual word. A word then has variable meaning owing to its immediate context of other words and possibly also a preference of certain contexts in its distribution. As a result, the frequency-based approach accepts all types of word-combinations as part of the field. (Granger and Paquot, 2008: 31)

In terms of morphology, the issue complicating the relationship between the two approaches is the question of what should be considered a word, as the central definition in the field of phraseology is that "phraseological units are made up of at least two words"

(Granger and Paquot, 2008: 32). The traditional phraseological approach tends to either completely omit compounds or consider only compounds fitting certain criteria that possibly point towards a nature other than that of a single unit such as stress. The frequency-based approach makes place for any orthographic word and therefore only excludes solid compounds. The issue of compounds and their place amongst the various types of words or units of language has been and is still being considered in the field of lexicology and lexicography and the varied views tend to give researchers and theorists leeway to arbitrarily set up criteria according to their own judgement as to which types of units they include or exclude from their research into word-combinations.

The relationship between phraseology and syntax is fairly complicated. In general, the traditional phraseological view tends to distance itself from syntax; just as compounds are left to the study of lexicon, grammatical considerations belong to syntax. The frequency-based approach, however, tends to include grammatical structures of word-combinations in its research, studying grammar-based combinations such as variable idioms[1] or pos-grams[2]. (Granger and Paquot 2008: 34)

Where phraseology overlaps with the field of discourse, the two approaches also tend to disagree. The traditional phraseological approach tends to take the interactional view of word-combinations focusing mainly on fixed units which serve certain pragmatic and discourse-organising functions. The frequency-based approach certainly takes these into account as well; it does, however, often attribute equal or even greater importance to text structuring word-combinations of varied size (from complete utterances to short snatches of words) which show mostly semantic and syntactic regularity. (Granger and Paquot, 2008: 35) As a consequence, the field of phraseology extends beyond fairly fixed pragmatic units and uncovers "a large stock of recurrent word-combinations that are seldom completely fixed but can be described as 'preferred' ways of saying things" and which are conventionalised and routine in language production. (Altenberg, 1998: 121, 122)

2.2.1.1 Categories of word-combinations

Various typologies of word-combinations have been established over the years, differing mainly because of the features that are used to categorise the units and their

[1] Variable idioms are constructions where at least one positions is variable in the specific form of the word but identifiable as a specific grammatical category (e.g. X *think nothing of* Vgerund) (Granger and Paquot, 2008: 32).
[2]Pos-grams are constructions occurring as strings of part of speech categories. (Stubbs, 2007: 91)

prioritisation. Granger and Paquot (2008: 35) establish that most typologies give prominence to one or more of the following features of phrasemes:

(1) internal structure (eg. verb + noun);
(2) extent: phrase- vs. sentence-level;
(3) degree of semantic (non-)compositionality;
(4) degree of syntactic flexibility and collocability;
(5) discourse function.

Typologies available in literature are often rooted in one of three areas – they are either designed for lexicological or lexicographic purposes, they are pedagogically-oriented or they take a psycho-linguistic perspective. Granger and Paquot (2008) present some influential typologies rooted in English lexicology and lexicography and also propose a categorisation emerging from the frequency-based approach of their own, as none have emerged from existing studies.

One of the most influential typologies reflecting the traditional phraseological approach is Cowie's categorisation of word-combinations into two main categories of composites, functioning below the level of a sentence, and formulae, functioning pragmatically as autonomous utterances (see Figure 2).

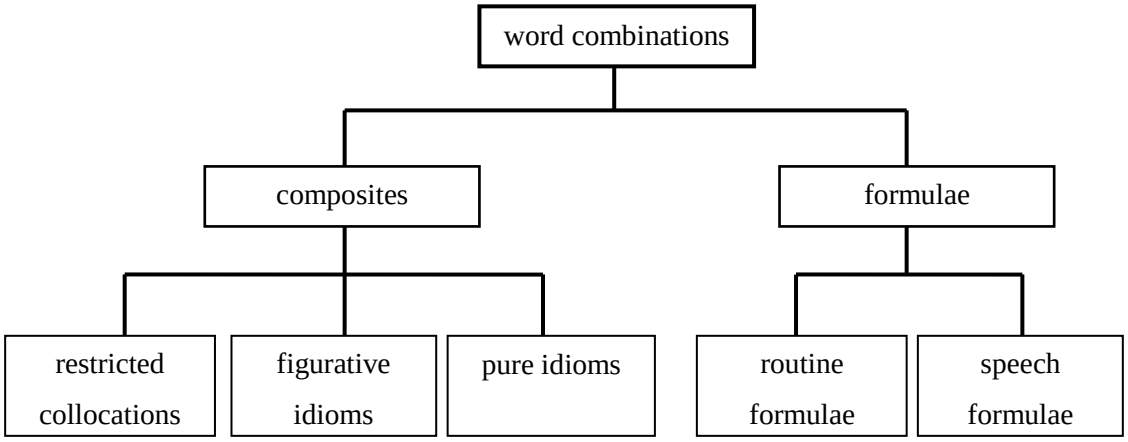

Figure 2: Cowie's (1988, 2001) classification of word combinations (in Granger and Paquot, 2008)

The subdivision of composites into restricted, *figurative* and *pure idioms* clearly shows the cline-like nature of the phraseological continuum, from the most variable and semantically transparent units at one end, to the most fixed and opaque on the other. The category of *restricted collocations* which is often referred to simply as "collocations" is characterised by

the units' restricted collocability and either figurative or specialised meaning of one of the parts of the collocation. Figurative and pure idioms differ in that the former hold a figurative meaning but also preserve a literal interpretation (e.g. *do a U-turn*), whereas the latter are semantically non-compositional and resist substitution of their elements. *Formulae* are seen as "sentence-like" units which "function pragmatically as sayings, catchphrases, and conversational formulae" (Cowie 1998b:4 / Granger and Paquot, 2008: 36). Cowie further subdivides this category into *routine formulae* (e.g. *good morning*), performing speech-act functions and *speech formulae* which are used to "organize messages and indicate speakers' or writers' attitudes," (Granger and Paquot, 2008: 36) e.g. *you know what I mean, are you with me?*

Another influential model of word-combination is proposed by Mel'čuk within his meaning-text theory, summarised by Granger and Paquot (2008) as follows: it corresponds very closely to Cowie's typology, using slightly different terminology. The two main categories of *semantic phrasemes* and *pragmatic phrasemes* roughly correspond, respectively, to Cowie's categories of composites and formulae. The important aspect of Mel'čuk's model (see Figure 3) is its treatment of collocations wherein he attempts to describe lexical preferences by way of lexical functions – these are general and abstract meanings which can be variously expressed depending on the specific lexical units to which it is applied. (Granger and Paquot, 2008: 37)

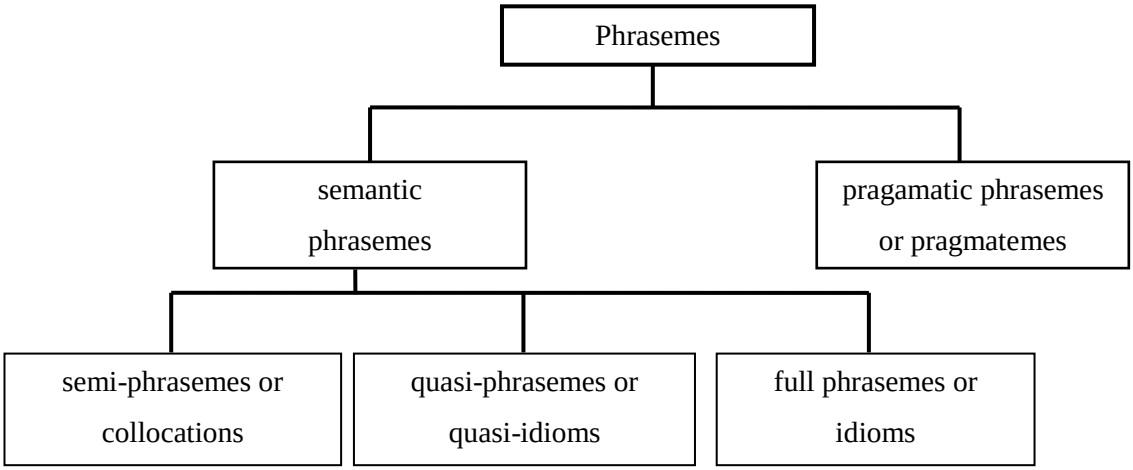

Figure 3: Mel'čuk's typology of word-combinations (1998, in Granger and Paquot, 2008)

The third and last typology is unique in that unlike Cowie's and Mer'čuk's models, it primarily focuses on the function of phraseological units in discourse. Burger's typology, shown in Figure 4, distinguishes at the top between three functional categories: referential, structural and communicative units. *Referential units* are further divided into two categories

according to a syntactic-semantic criterion. *Nominative* phraseological units are part of a sentence and refer to an object, a phenomenon or a fact of life. This category again roughly corresponds to composites. The traditional approach shows in the division of the nominative units into collocations, partial idioms and idioms according to their variability and semantic transparency. Propositional phraseological units more often than not function at the sentence-level, but sometimes function at the level of text; they involve proverbs and idiomatic phrases corresponding to Cowie's and Mel'čuk's categories of formulae or pragmatic phrasemes. Communicative units, corresponding to Cowie's routine formulae, fulfil an interactive function in discourse and help textually organise interaction. The last category of structural phraseological units includes combinations which establish grammatical relations and is mostly considered as having the least interest to phraseology by Burger himself. (Granger and Paquot, 2008: 38)

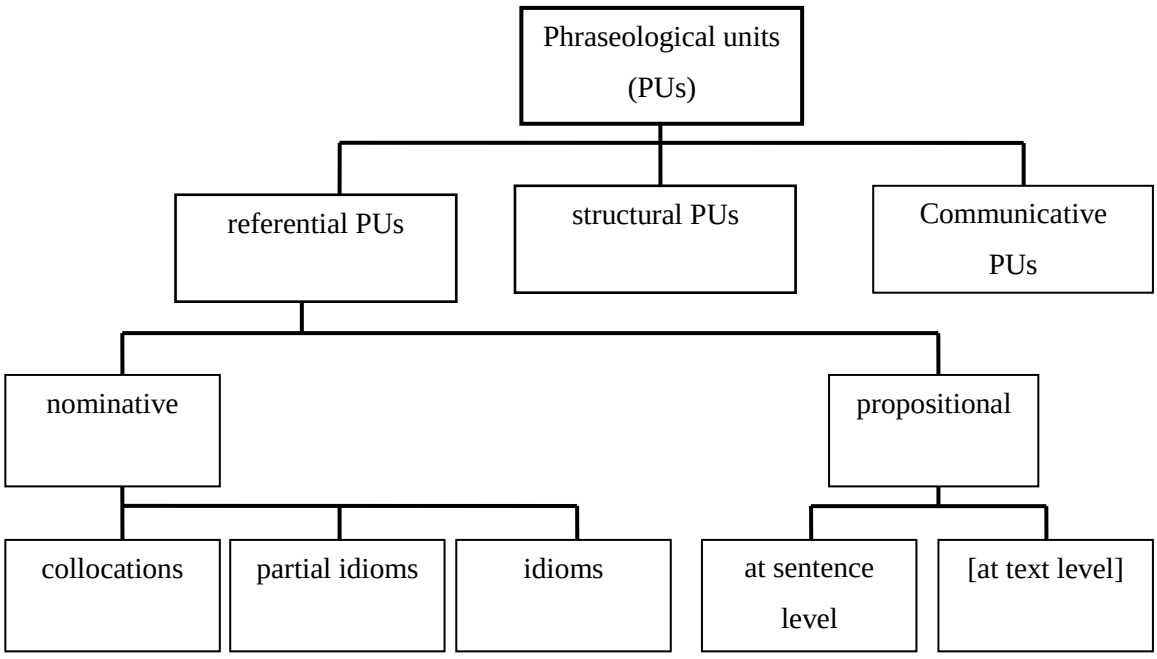

Figure 4: Burger's typology of word-combinations (1998, in Granger and Paquot, 2008)

Granger and Paquot further introduce their own frequency-based typology. It is precisely this categorisation that seems to be most relevant to the present study, as the data involved in the research are extracted based on frequency criteria. It also serves as a useful introduction into the topic of recurrent word-combinations and so the classification is included at the beginning of chapter 2.2.2. The last note to be made in the present chapter

should be given to Granger and Paquot's view as to what terminology is suitable for studies based on automated, assuming also frequency-based, extraction of research data.

> To refer to the results of automated extraction, we advocate the use of the terms in Figure [5]. This means that in our view the term 'collocation' should not be used to refer to statistical word co-occurrences but instead kept in its traditional meaning of usage-based lexically restricted combination. (Granger and Paquot, 2008: 42)

2.2.2 Recurrent word-combinations: views, terminology and research

As has been stated before, there is no categorisation of word-combinations emerging from the studies rooted in the frequency-based approach to phraseology. Granger and Paquot do, however, draw up a typology built on the units that have appeared in research based on two types of extraction. Figure 5 reflects this typology, following the division of research base on n-gram analysis and research based on co-occurrence analysis.

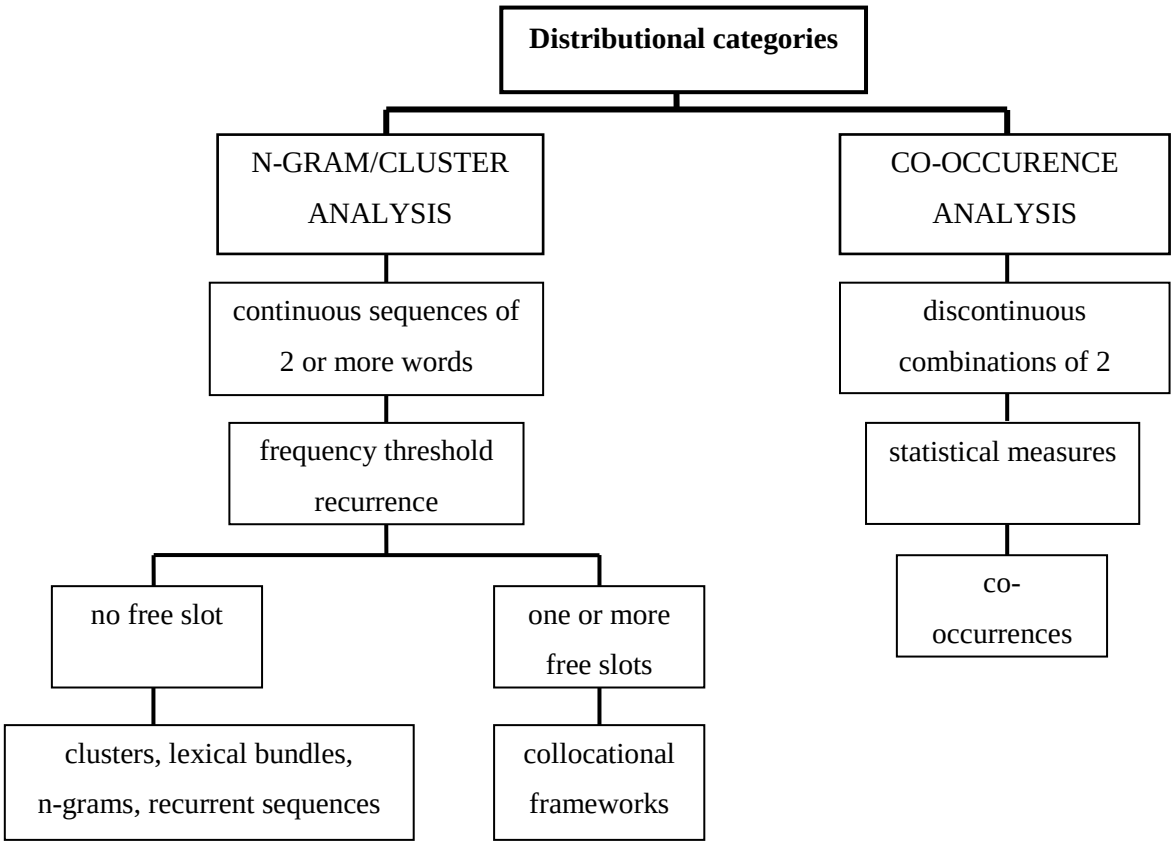

Figure 5: Distributional categories of word-combinations (Granger and Paquot, 2008)

Co-occurrence analysis may be defined as the statistical uncovering of significant word co-occurrences. Units retrieved in this kind of analysis are usually referred to as collocations or as collocates and the definitions of these can be varied, depending on how far a researcher is willing to extend the range of co-occurrence in relation to the headword. As has been stated before, some researchers, including Granger and Paquot, prefer to refer to these as co-occurrence or co-occurrent when using statistical methods of analysis.

N-gram analysis is in extraction method which helps extract recurrent continuous sequences of two or more words. The terminology for the units extracted by this method tends to differ study from study. As has already become clear, the frequency-based approach tends to be "much less preoccupied with distinguishing between different linguistic categories and subcategories of word combinations or more generally setting clear boundaries to phraseology." (Granger and Paquot, 2008: 29) As a result, terminology and criteria have been used and created ad hoc, depending on what purpose a study is supposed to serve. Each study tends to follow its own artificial criteria of what should be considered a unit or not.

Some of the terms used in studies using frequency-based methods for research are as follows: n-grams, lexical bundles, clusters, chains, recurrent word-sequences and recurrent word-combinations. It should be mentioned that certain articles and studies may use the n-gram as a cover term for the above mentioned terms. (Stubbs: 2007) The following paragraphs attempt to outline a general idea of the variation involved in these studies. After providing a short overview of some of them, it should also become clear which paths the present research might follow.

Most studies are rooted in the overarching idea of the prevalence of formulaic language, a language of prefabricated sequences in short, as a reflection of language processing and language storage. Sinclair's idiom principle has already been mentioned. Further mention can be made for psycholinguistic research done by Allison Wray (2002) which "lends support to the idea that human language production and processing includes multi-word sequences as single units." She argues that while the proposed claim in her studies may be the dual system of analytical and holistic processing of language, the latter is much preferred because it does not require so much effort. Nevertheless, studies tend to vary exceptionally in how they choose to identify and study word-combinations and in the terminology they use. "Given the variety of purposes in studies of multi-word sequences, it is not surprising that the sequences have been identified in diverse ways." (Conrad and Biber, 2005: 58):

Overall, six characteristics tend to be singled out as most important (Conrad and Biber, 2005: 57):

- fixedness
- idiomaticity
- frequency
- length of sequence
- completeness in syntax
- semantics or pragmatics
- intuitive recognition by speakers in the native community

Studies of pure idioms, to use Cowie's terminology, such as *kick the bucket* give priority to fixedness, idiomaticity, completeness and intuitive recognition by native speakers. Conrad and Biber (2005: 57) point out that these criteria, while valid, tend to restrict the playing field to a very small portion of the stock of word-combinations and also take time to mention that constructions such as pure idioms are rarely attested in normal use and/or face-to-face conversation.

Studies of collocations, on the other hand, prioritise frequency and two-word relationships. The studies are also based on an artificially chosen threshold for the strength of collocation as the authors deem appropriate; meaning intervening words between the two elements of the collocate are allowed. Fixedness of form is therefore not an issue, neither is identification by native speakers. Idiomaticity also does not appear as a criterion, whereas semantic unity of the collocation is.

The study of Nattinger and Decarrico (1992) focuses on *lexical phrases* and the authors give primacy to semantic or pragmatic completeness, intuitive recognition and fixedness. Without considering frequency, they provide a typology of these phrases in categories such as topic markers or sentence builders. While not accounting for the rate of occurrence in natural discourse, the study provides useful data for pedagogical purposes.

Another study based on the n-gram analysis method assumes the quintessential bottom-up approach of frequency-based phraseology. Altenberg (1998) calls phraseology a "fuzzy part of language" and in a fitting manner assumes this very "non-committal approach" to its study. His work brings to the field of phraseology the term *recurrent word-combination*. It is simply defined as "any continuous string of words occurring more than once in identical form." (Altenberg 1998: 101) Altenberg's (1998) research is only restricted for size and

range of occurrence (3-grams occurring at least ten times in the corpus) for the purpose of making the study manageable. He admits that these criteria are largely arbitrary and do not really speak to their phraseological status, but also assumes that the frequency threshold speaks to the prevalence of the extracted combinations in language usage. This specific study also excludes any combinations which include unintentional repetition or stuttering (e.g. *the the the, I was I was*). This is especially relevant, as the data used for the present study is certain to show such combinations in the process of automatic extraction. Altenberg (1998) groups his extracted combinations by their syntactic status (e.g. *full clauses, stems*) and then considers functions of the word-combinations in discourse and successfully uses his findings to support the theory of processing where the open-choice principle alternates with the idiom-principle, the latter of which is considered dominant.

In their major corpus-based work dealing with recurrent word-combinations (LGSWE) Biber et al. distinguish between idioms, collocations, lexico-grammatical associations and *lexical bundles***.** The term lexical bundle is, in short, introduced as a kind of "extended collocation." They follow Altenberg's frequency-driven, fixed-word approach; the focus is on 3- and 4-grams which span the range of at least 5 texts. This, admittedly again arbitrary, choice of settings has proven to show results over the course of many studies (eg. Götz, 2013; Conrad and Biber, 2005). In LGSWE, they "emphasized the structures of lexical bundles, and discussed the structures' associations with various discourse functions." (Conrad and Biber, 2005: 59) In their later study (2005), they adopt the very same approach of register perspective – here conversation and academic prose, further extending their research into discourse functions and presenting a preliminary classification of lexical bundles into functional categories (see more in chapter 2.2.3).

2.2.2.1 Recurrent word-combinations: formulaic sequences and previous contrastive studies of learner language

Research into the use of recurrent word-combinations in learner language is a rather recent endeavour, but a fairly large body of studies has been written, be it studies solely focused on the use of formulaic language or the use of formulaic language analysis as part of larger research into fluency. This chapter introduces the topic of formulaic sequences as treated in relevant research and gives an overview of previous contrastive studies of recurrent word-combinations in non-native speech.

As has already been mentioned, there are two general approaches to the study of formulaic language: phraseological and frequency-based or distributional. (Granger &

Paquot, 2008: 27) Defining generally applicable criteria for whether a sequence is formulaic or not is therefore an ongoing issue. Ellis (2012), for example, introduces a fairly clear-cut set of criteria by which he defines formulaic language and which can be traced in many studies which deal with formulaicity of recurrent word-combinations. These three main criteria are frequency, association and native norms.

The criterion of **frequency** is fairly self-explanatory. Ellis, however, makes an important observation; that not every frequently recurrent sequence is formulaic and not every formula is necessarily frequent. (Ellis, 2012: 27) This is the reason Ellis adds the other two criteria. **Association** puts focus on the strength of co-occurrence of words in a sequence rather than on raw frequency. And finally, **native norms** are an additional criterion of formulaicity that reflects native-like selection and native-like fluency. (Ellis, 2012: 29) It should be mentioned that all three of these are part of the six most important characteristics of word-combinations (Conrad and Biber, 2005: 57) listed in chapter 2.2.2, which highlights the undeniable part formulaicity holds in research on word-combinations.

While these criteria may be, perhaps inevitably, given unequal importance in most studies, researchers often do make use of more than one of them. They also often play an important role during the data-collecting phase, namely in the extraction of relevant data from language corpora. The present study's focus on recurrent word-combinations results in a certain bias towards the frequency-based studies. Research focused on formulaic sequences often makes use of extraction methods which retrieve all repeated sequences fully automatically; this method has been called "an illustration of corpus linguistic methodology at its most heuristic, i.e., as a raw discovery procedure." (De Cock, 2004: 227) Some studies then focus simply on the frequency of recurrence and do not go further to establish which combinations are actually formulaic and which are not.

Association, on the other hand, often gains importance in research focused on collocations. A fairly common approach to measuring the strength of association in collocates in learner language "is to make use of association measures computed from reference native or expert corpora, rank co-occurrences in learner corpora on that basis, and judge their acceptability." (Paquot and Granger, 2012: 135) Although association measures tend to become unreliable with low-frequency data, they need not lose their function needlessly; their usefulness becomes obvious in studies focused on naturally frequent phenomena or, as the quotation above suggests, a comparison of learner-language data with native-speaker data. What does not escape notice is that even as they outline approaches centring on association, Paquot and Granger (2012: 135) mention native norms.

The importance of Ellis' native norms criterion is often taken into consideration, be it in contrastive studies of native and non-native language production or simply in the analysis of non-native production. Götz (2013) and Wood (2010), for example, both include native norms as a criterion in their research. Wood uses a fairly complex set of criteria to determine fomulaicity, but native norms are given a lot of weight nevertheless. Wood consults informed native-speaker judgement to determine whether certain sequences are formulaic or not; the judgement is informed in a sense that the speakers had read literature on the topic of formulaicity and were given certain criteria to judge by. Götz, on the other hand, observes the native norm in a way that seems to be in line with Paquot and Granger's suggestion. While making use of Altenberg's frequency-driven approach (see chapter 2.2.2), Götz uses in her contrastive study of native-speaker and a non-native-speaker corpora the total set of sequences extracted from the native-speaker corpus as a reference list for the judgment of formulaicity of sequences used by speakers of both corpora; that is, she "compared each of the speakers individual performances with the occurrences in the LOCNEC [native speaker] totals (...) and included these 3-grams and 4-grams as a formulaic sequence for a speaker only if they also occurred in the LOCNEC [native speaker] total." (Götz, 2013: 102)

A fuller and more specific set of criteria to determine whether a sequence, no matter if it is frequent or not, is formulaic is outlined by Wray. (2002: 31-43) Structure and form, compositionality, fixedness (or semi-fixedness) and the phonological form. The latter criterion especially puts focus on the pre-fabricated nature of word-sequences, as it is seen to reflect speakers' processing more immediately, and it can, and possibly should, play an important part in analysing spoken language. Wray mentions, for example, a lack of pauses and hesitation phenomena, precision and speed of articulation, intonation pattern and stress.

As the present study centres around the contrastive analysis of spoken data, the following few paragraphs shortly introduce previous research into recurrent word-combinations in the speech of native and non-native speakers. Most studies make use of Altenberg's method of extraction (De Cock, 1998; De Cock, 2004; Götz, 2013; Larsson Aas, 2011), with varying strictness of frequency and length-criteria. All three of these studies centre around the contrastive analysis of comparable native-speaker and non-native-speaker corpora and all three use the same native-speaker corpus which is the source of native spoken data in this thesis as well. De Cock (2004) and Larsson Aas are very similar in the breadth and nature of their analysis, in that they do not confine themselves to a specific sequence length and perform rather extensive qualitative analysis of their data, whereas Götz focuses

on four- and three-word sequences and generally does not explore the data from a qualitative point of view. From a quantitative standpoint De Cock's (2004: 231) findings concerning French speakers, for example, show that native speakers tend to have a more varied supply of recurrent word-combinations and they also use them more frequently than non-native speakers. This is also true for Götz's (2013: 102) findings in her study of fluency concerning German speakers. Larsson Aas's research, in addition, shows that highly recurrent word-combinations in the Swedish and Norwegian corpora show salient similarities with those that are highly recurrent in the native-speaker corpus.

De Cock's (2004) and Larsson Aas also discuss overuse and underuse of certain combinations (e.g. *I don't know* for overuse; *that's right* and certain vagueness devices for underuse) and they both analyse the pragmatic/discourse functions of certain combinations. De Cock's research also includes hesitation items (filled pauses) and/or repeats. Frequency results before the exclusion of combinations containing these phenomena actually seemed to suggest that non-native speakers use more recurrent word-combinations than native speakers. Larsson Aas, following De Cock's lead in her analysis, also includes filled pauses and repeats, as they "may perform important functions in spoken discourse." (Larsson Aas, 2011: 116) The inclusion seems to be useful in that during the analysis it was, for example, shown that Norwegian and Swedish speakers used fewer filled pauses and repeats than French speakers in De Cock's previous research. Larsson Aas claims that it "is possible to hypothesize that if the usage patterns of pauses and repetitions in LINDSEI-SW and LINDSEI-NO are more similar to those found in native English speech, this is a reflection of a higher level of general proficiency among these learners." (Larsson Aas, 2011: 67)

In the area of formulaicity, these studies differ. De Cock's older study (1998) presents a manual filtering process which reduces the automatically extracted list of recurrent word-combinations to a list of actual formulaic usage, whereas the newer study (2004) does not mention the issue of formulaicity at all. Larsson Aas dedicates a section of her thesis to ruminate on the possible implications her findings might have in the context of formulaic language and Götz observes the previously mentioned native norms criterion to slightly improve the chances of the combinations used in her quantitative analysis actually being formulaic.

2.2.3 Functional categorisation of recurrent word-combinations

The efforts put into creating a functional framework for recurrent word-combinations are ongoing. That is not to say that the function of recurrent word-combinations is not taken into consideration in many studies; it is, however, made on a more case-to-case basis rather than as part of an effort to create a broader, more general categorisation. This chapter introduces two prominent studies which have managed to reach certain conclusions and create their own functional categories.

Function	**Effects**	**Type**	**Examples**
Processing short-cuts	Increased production speed and/or fluency	Standard phrases (with or without gaps)	Put the kettle on, will you?
		Standard ideational labels with agreed meanings	Personal computer; bullet point; the current economic climate
Time-buyers	Vehicles for fluency, rhythm and emphasis	Standard phrases with simple meanings	Make a decision; draw a conclusion; one way or another
	Planning time without losing the turn	Fillers	If you want my opinion; if you like
		Turn-holders	And another thing
		Discourse shape markers	Firstly…; secondly…
		Repetitions of preceding input	(A: What's the capital of Peru?) B: What's the capital of Peru? (Lima, isn't it?)
Manipulation of information	Gaining and retaining access to information otherwise unlikely to be remembered	Mnemonics	Thirty days hath September…
		Lengthy texts one is required to learn	Shall I compare thee to a summer's day?
		Rehearsal	*Rehearsing a telephone number while looking for a pen*

Table 1: Formulaic sequences as compensatory devices for memory limitations (Wray and Perkins, 2000: 16)

Two studies by Wray (Wray and Perkins, 2000; Wray, 2002) divide formulaic language into two major groups according to their function: categories of formulaicity as a *short-cut in processing* (table 1) and formulaicity as a *tool for social interaction* (table 2).

The tables are reproduced from the older study, with a few examples removed for the sake of space. Wray later (2002) abandons the categorisation in the processing group (table 1) and also disclaims the possibility of these categories becoming a fixed system (2002: 70):

> It is not intended to imply a typological match. It is an organization of convenience, which will suitably reveal patterns in the data. Inevitably, some types of sequence will turn up in more than one category, but this is simply an indication of the kind of complexity that we are dealing with.

Function	**Effects**	**Type**	**Examples**
Manipulation of others	Satisfying physical, emotional and cognitive needs	Commands	Keep off the grass; hand it over
		Requests	Could you repeat that please?
		Politeness markers	I wonder if you'd mind…
		Bargains, etc.	I'll give you __ for it.
Asserting separate identity	Being taken seriously	Story-telling	You're never going to believe this, but…
		Turn claimers and holders, etc.	Yes, but the thing is…; Thank you very much (in response to invitation to speak)
	Separating from the crowd	Personal turns of phrase	I wanna tell you a story (*Max Bygraves*)
Asserting group identity	Overall membership	'In' phrases	Praise the Lord!; as the actress said to the bishop
		Group chants	We are the champions
		Institutionalized forms of words, etc.	Happy birthday; dearly beloved
		Rituals	Our Father, which art in heaven
	Place in hierarchy (affirming and adjusting)	Threats	I wouldn't do that if I were you
		Quotation	"I wouldn't want to belong to any club that would have be as a member" (*Groucho Marx*)
		Forms of address	Your Highness
		Hedges, etc.	Well I'm not sure (polite denial)

Table 2: Formulaic sequences as devices of social interaction (Wray and Perkins, 2000: 14)

	Conversation	Academic prose
I Stance Expressions		
I-A. Epistemic Stance		
Personal:	I don't know what; I think it was; you know what I	
Impersonal:		the fact that the
I-B. Attitudinal/Modality Stance		
• Desire	I don't want to; if you want to; I would like to	
• Obligation/Directive		
Personal:	you don't have to; you want me to	
Impersonal:		it is necessary to
• Intention/Prediction		
Personal:	I was going to;	
Impersonal:		it's going to be;
• Ability		
Impersonal:		it is possible to
II. Discourse Organizers		
II-A. Topic Introduction/Focus	what do you think; have a look at; do you know what	
II-B. Topic Elaboration/Clarification	nothing to do with; was going to say; what do you mean	on the other hand
III. Referential Expressions		
III-A. Identification/Focus		one of the most
III-B. Imprecision	or something like that	
III-C. Specification of Attributes		
• Quantity Specification		per cent of the
• Tangible Framing Attributes	in the form of	
• Intangible Framing Attributes		in the case of; the way in which
III-D. Time/Place/Text Reference		
• Time Reference		at the same time; at the time of
• Multi-Functional Reference	the end of the	at the end of
IV. Special Conversational Functions		
• Politeness	thank you very much	
• Simple Inquiry	what are you doing	
• Reporting	I said to him	

Table 3: Functional Classification of 4-word lexical bundles with frequencies over 40/million words (Conrad and Biber, 2005: 65-66)

Another set of categories, created while using actual data from the *Longman Spoken and Written English Corpus*, is presented by Conrad and Biber (2005). They use a list of common 4-word-combinations and sort them into four overarching categories which are shown in table 3, which was reproduced with occasional deletion of some examples, again, for the sake of space.

Ädel and Erman in their study of recurrent word-combinations express several reservations (2011: 88) about the functional classification used in the research of Chen and Baker (2010). This classification closely corresponds to that of Conrad and Biber (2005) and the reservations fully apply to both classifications included in the chapter. One of the reservations is that the classifications remain preliminary and more or less unchanged and the main problem is that no clear criteria are established for what makes a word-combination part of a certain category or subcategory. Some subcategories are intuitive, but others are vague and lead to inconsistencies in research. Another problem, inherent to the practice of functional categorisation, is the multifunctionality of many word-combinations. "It is therefore necessary to consider the extended context to determine what the predominant function is." (Ädel and Erman, 2011: 88)

3 MATERIAL AND METHOD

The data used in the present study was extracted from two comparable corpora of spoken English. Both are part of the Louvain International Database of Spoken English Interlanguage (LINDSEI) project, which was launched in 1995 as a spoken counterpart to the International Corpus of Learner English project (ICLE), which is a learner corpus of written English containing argumentative essays written by higher-intermediate to advanced learners of English from several mother tongue backgrounds. LINDSEI is then a spoken corpus containing oral data produced by advanced learners of English also from several mother tongue backgrounds. The individual subcorpora have presented a possibility to compare spoken learner language of speakers of various L1; completed subcorpora have been compiled for Bulgarian, Chinese, Czech, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, Swedish, Taiwanese and Turkish speakers. In addition, a comparable corpus of spoken native English has been compiled as the Louvain Corpus of Native English Conversation (LOCNEC), making comparison possible also between learner language and native production.

Both corpora used in this study (the Czech subcorpus of LINDSEI and LOCNEC) include data from 50 speakers. The data was acquired through informal interviews. All interviews last for approximately 15 minutes and follow the same pattern of three separate tasks. For Task 1 the students speak on a chosen topic, Task 2 involves a conversation on common topics such as personal interests, studies etc. and for Task 3 students are asked to use four pictures to reconstruct a story. Students are given time to prepare for Task 1, although they are not permitted to write notes, whereas they have no time to prepare for Task 2 and 3. The Czech subcorpus (LINDSEI_CZ) is made up of 95,904 word-tokens[3] of interviewee speech and LOCNEC is made up of 117,417 word-tokens.

The choice was made, however, to include only Tasks 1 and 2 in this particular study. The reason is twofold: firstly, Task 3 is arguably more restricted in the language choices the students make, as they are guided by the same set of pictures to tell a story, and it is significantly less interactive than the other two tasks. Secondly, the exclusion of Task 3, in which the students produced less language than in the other tasks, presented an opportunity to reduce the size of the data, which would otherwise have been too large in its entirety for

[3] Not to be confused with 4-gram tokens defined later in the chapter (footnote number 6). What is here called a word-token is an occurrence of an orthographic word in the corpus.

this particular study. The size of the corpora is then reduced to 83,434 word-tokens for LINDSEI_CZ and to 114,768 for LOCNEC.

The method used to investigate recurrent word-combinations in the speech of native speakers and Czech speakers of English is the corpus driven "recurrent word-combination" method used in studies by for example Altenberg (1998), De Cock (2004) or Götz (2013). This method involves automatic extraction of word-combinations using a specialised software, making the analysis a kind of raw discovery procedure which does not presuppose any linguistic categories or pre-established lists of word-combinations. "The results yielded by the automatic extraction are a useful and powerful starting point as they arguably lead the researcher to take into consideration a series of frequently used clusters he or she may otherwise have overlooked because of their lack of psychological salience." (De Cock, 2004: 227)

The present study focuses on recurrent 4-word-combinations (also referred to as 4-grams).[4] The data is analysed in two steps, first quantitatively and then qualitatively. The initial quantitative analysis compares results of both corpora and consequently also compares them to previous research involving comparison of native-speaker and learner language production (specifically De Cock, 2004). The recurrent word-combinations presented in this study for both the quantitative and qualitative analysis were extracted using the AntConc 3.4.4w software[5].

For the quantitative part of the analysis, the frequency threshold was set at 7 occurrences in the corpus across the range of at least 5 speakers. The minimal frequency and minimal range were both chosen artificially with the goal of extracting an amount of data adequate to the ambitions of this study. The data extracted by the program was then modified after extraction and before the actual analysis in the following ways. Firstly, the program included in the extracted data 4-grams which went over the speaker turn i.e. 1-3 words contained in the specific instance of a 4-gram occurred at the end of one speaker-turn and the following 1-3 words of the 4-gram started at the beginning of the following speaker-turn; these instances were therefore manually removed from the data as they do not present actual continuous word-combinations. In cases where the removal of these occurrences caused the specific type[6] of 4-gram to fall below the set frequency or range threshold, the whole type

[4] Contractions (e.g *don't, I'd*) are accepted as one word, therefore combinations such as *I don't know if* are considered to be 4-grams.
[5] Available at http://www.laurenceanthony.net/software/antconc/
[6] Each different word-combination is considered a different **type** and each occurrence of a word-combination a different **token**.

was removed from the excerpted data. All statistical significance of frequency/occurrence findings in the study are determined using the log-likelihood test[7].

The more qualitative portion of the analysis required a slightly different approach to the material, the reason again being the size of the data that could be feasibly analysed excerpt by excerpt in the present study. The qualitative analysis is then performed on a smaller sample of speakers from both corpora. The criterion for the choice of individual speakers was speech rate, presenting a possibility to investigate the link between speech rate and use of recurrent word-combinations. 15 speakers were chosen from each corpus according to speech rate data provided in Gráf (2015), which can be observed in Figures 6, 7, 8 and 9, also lifted directly from Gráf (2015: 131-133). Both samples consist of: 5 of some of the slowest speakers, 5 of some of the speakers with an average rate of speech and 5 of some of the fastest speakers. The speakers chosen for the LINDSEI_CZ sample were then speakers 24, 41, 20, 42, 34, 36, 14, 17, 46, 33, 11, 12, 26, 9 and 48. For the LOCNEC sample it was speakers 4, 41, 34, 42, 31, 40, 17, 13, 29, 27,12, 48, 9, 49 and 3.

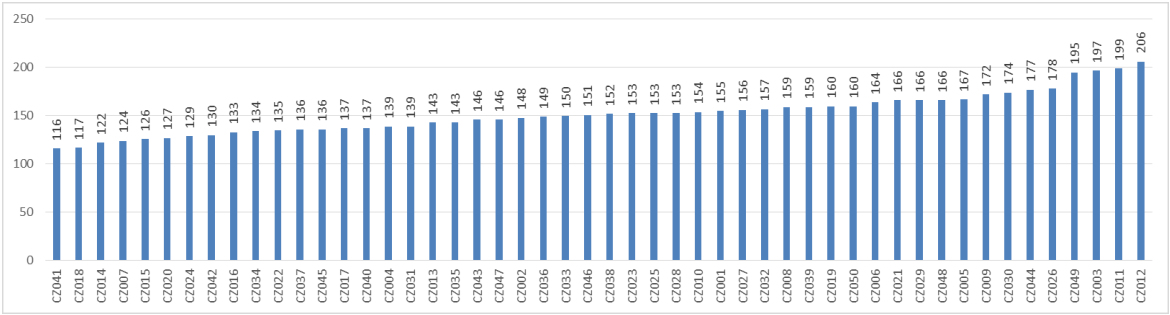

Figure 6: Non-native speech rates in Task 1 for all LINDSEI_CZ speakers (figures above bars represent words per minute)

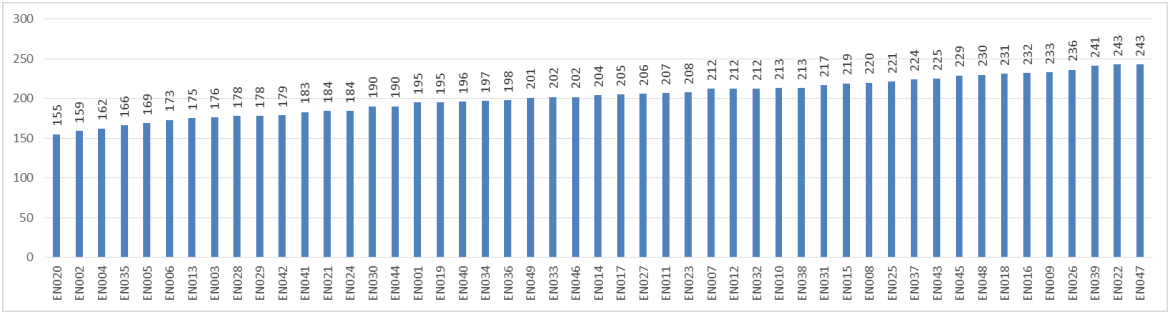

Figure 7: Non-native speech rates in Task 2 for all LINDSEI_CZ speakers (figures above bars represent words per minute)

[7] The *Log-likelihood and effect size calculator* was used. (http://ucrel.lancs.ac.uk/llwizard.html)

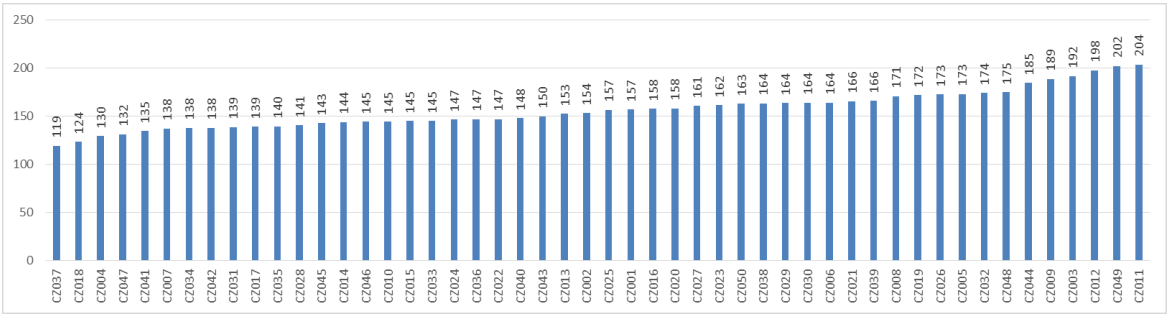

Figure 8: Native speech rates in Task 1 for all LOCNEC speakers (figures above bars represent words per minute)

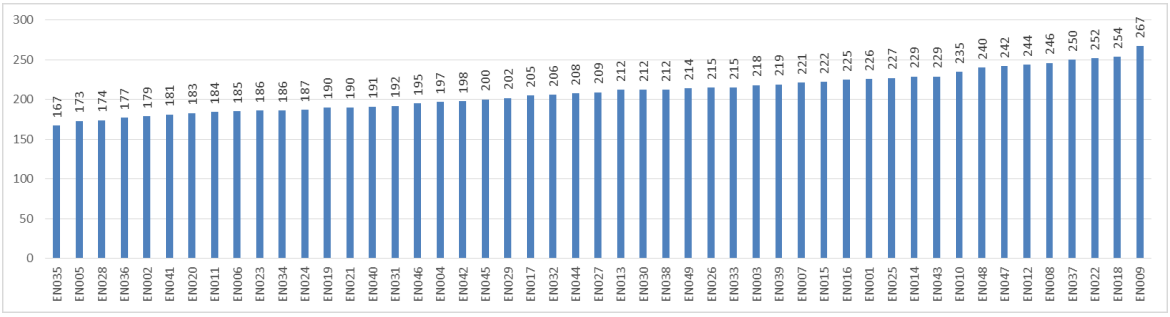

Figure 9: Native speech rates in Task 2 for all LOCNEC speakers (figures above bars represent words per minute)

The frequency threshold for the qualitative analysis was set at 2 occurrences and the range of at least 2 speakers. The lower thresholds were chosen due to the fact that such a small sample would otherwise not have yielded enough 4-gram types. The low frequency threshold admittedly invalidates the possible relevance of the study to research on "recurrent" word-combinations, as 2 occurrences can hardly be considered as proof of recurrence. This was partly remedied by a deliberate step in the analysis. The types which appeared during the extraction from the smaller sample of speakers were compared to the types acquired during the extraction from Tasks 1 and 2 of the LINDSEI_CZ and LOCNEC corpora (see Appendix 3); only the types which appear in the respective lists for the quantitative analysis were included in the qualitative analysis. Speakers from the LINDSEI_CZ sample shared 29 types with the list of types extracted during the quantitative analysis. Because the LOCNEC sample yielded more shared types, the list was reduced to 29 most frequent shared types as well. (see Appendix 4) The data extracted for the qualitative analysis was also modified in much the same way as the data involved in the quantitative analysis – 4-grams crossing the turn border were removed. Furthermore, occurrences of 4-grams which contained unintentional repeats or hesitation items such as filled pauses (e.g. *um*, *er*) or truncated words

(e.g. *I d= I don't*) were also removed, following the lead of Altenberg's study (1998) and taking into consideration the importance of recurrent word-combinations in the study of fluency; hesitation items disrupt fluency, whereas repeats support fluency in a much different manner to recurrent word-combinations and their use cannot be generalised even in the speech of a single individual, making them irrelevant to the goals of the present study.

The analysis is conducted on excerpts from the speaker samples – the LINDSEI_CZ sample provided 111 excerpts for analysis, whereas the LOCNEC sample provided 133 excerpts (see Appendix 6). The study, however, also uses the quantitative results from the quantitative analysis and also other quantitative data, usually acquired through immediate searches, that provide opportunities to further the reach of the analysis. This way the conclusions made during the analysis have a wider validity, not only inside of the small speaker samples. This approach is outlined in more detail in Chapter 4.2.

An issue encountered during the extraction of all data which must be addressed was the presence of overlaps where specific occurrences of 4-grams appeared in two different types, i.e. they were most probably 5-word or even 6-word sequences. An example taken from the data extracted from LINDSEI_CZ of such an overlap is shown below. Type 1 4-gram occurs in the corpus with a certain frequency and range, whereas type 2 occurs with a different frequency and range.

Type number	**Type**	**Frequency**	**Range**
Type 1	I would like to	68	33
Type 2	like to talk about	8	7

If the occurrences of Type 2 which also appear for Type 1 are removed, the frequency is lowered (to 5 occurrences) and the range narrows (to 4 speakers).

Type number	**Type**	**Frequency**	**Range**
Type 1	I would like to	68	33
Type 2	like to talk about	5	4

This is obviously due to the fact that the 3 shared occurrences are of a sequence of 6 words - *I would like to talk about* - whereas the remaining 5 occurrences of *like to talk about* are not preceded by *I would* and are therefore not identical. This is certainly a setback of the extraction method, or the software used for the extraction. The difficulties and possible implications which these overlaps might have are only discussed in Chapter 5. The issue is

not otherwise addressed and is more or less accepted, for the length of the present study, as part of the imperfect method of extraction. Specifically, these overlaps only directly influence the data in the sample of speakers for the qualitative analysis and only in such a way that the list of all 4-gram occurrences provided for the analysis in the Appendix was modified with these overlaps in mind; the list provides each occurrence only once, not once for each type in which it occurs. This is for practical reasons – identification of occurrences used as examples in the analytical chapter is made easier. The list of types of 4-grams used by the smaller sample of speakers also provided in the Appendix was left as it was extracted as there is no feasible way to reflect the presence of these overlaps.

The **research questions** posed for the above outlined analysis are as follows:

1. Do the quantitative results support previous research on the differences between non-native speaker and native speaker production of recurrent word-combinations, i.e. do Czech speakers use a smaller number of types of these combinations and do they also use them less frequently?
2. Do repeats and/or hesitation items affect the results of the analysis?
3. What kinds of recurrent word-combinations do Czech and native speakers use? Are there significant differences between the two groups?
4. Can all recurrent word-combinations extracted from the data be considered 4-grams? What makes them so?

4 ANALYSIS

The investigation of recurrent 4-word-combinations is divided into two parts. Chapter 4.1 is dedicated to the quantitative analysis of the data from LINDSEI_CZ and LOCNEC. The steps in the analysis closely follow De Cock's study (2004), which investigates production of recurrent word-combinations of French speakers of English and native speakers and the results of her study are compared with the results in the present thesis. Chapter 4.1 also considers the issue of repeats and hesitation items. Chapter 4.2 is dedicated to the qualitative analysis of the speech of a smaller sample of speakers from both corpora.

4.1 Quantitative analysis of recurrent word-combinations in LINDSEI_CZ and LOCNEC

The quantitative analysis of the types and tokens used by Czech and native speakers of English is modelled in such a way that it is comparable with De Cock's study (2004) which compares French speakers to native speakers of English. De Cock uses her analysis to see whether the findings lend support to the hypothesis that learners use fewer prefabricated sections than native speakers, also admitting that the use of a sample of learners of a single L1 and of recurrent and continuous sequences (which need not necessarily be prefabricated) cannot fully investigate the validity of this hypothesis. The following findings are then compared with De Cock's findings rather than used to confirm or infirm this hypothesis.

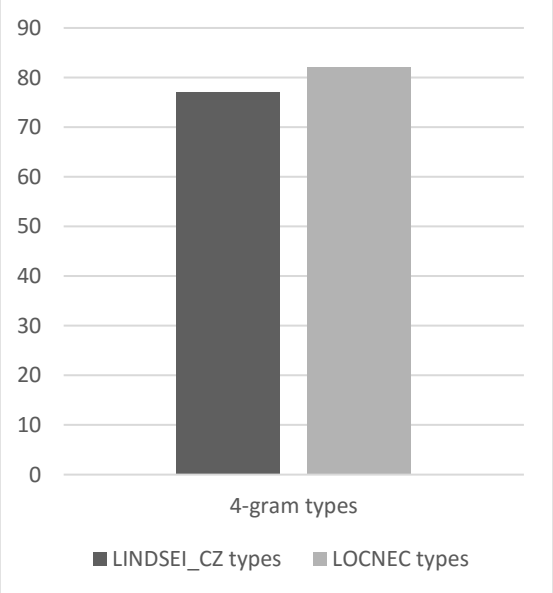

Figure 10: Czech vs native speaker 4-gram types

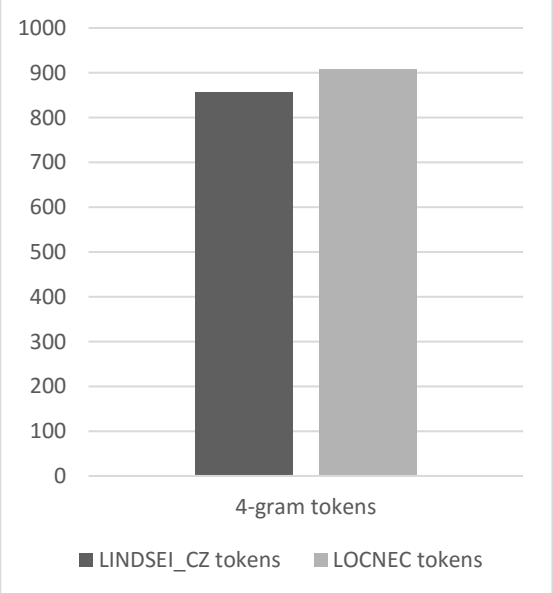

Figure 11: Czech vs native speaker 4-gram tokens

Figures 10 and 11 show the relative frequencies (occurrences per 100,000 word-tokens) of 4-grams in the production of both groups of speakers. While there is a slightly greater number of 4-gram types (81) and tokens (908) in LOCNEC than in LINDSEI_CZ (77 types and 856 tokens), the difference is not statistically significant. These findings differ from De Cock's study, where the French speakers use significantly more 4-gram types ($p \leq 0.05$) and tokens ($p \leq 0.005$)[8] than native speakers. The findings then show the Czech speakers to be more similar to native speakers than to the French speakers in their production of 4-grams.

De Cock's study also shows that a great number of the most frequent 3-gram types in the French subcorpus of LINDSEI contains repeats and/or hesitation items. This is the case also for the data extracted from LINDSEI_CZ and LOCNEC. Tables 4 and 5 show 20 most frequent 4-gram types appearing in LINDSEI_CZ and LOCNEC, respectively.

Rank	**4-gram**	**Rank**	**4-gram**
1	I would like to	**11**	at the same time
2	it was it was	**12**	er I think that
3	yeah yeah yeah yeah	**13**	decided to talk about
4	or something like that	**14**	but on the other
5	I think it was	**15**	I was I was
6	on the other hand	**16**	something like that and
7	and it was really	**17**	because I think that
8	I don't know I	**18**	I don't really know
9	I'm not really sure	**19**	I have to say
10	in the Czech Republic	**20**	I would really like

Table 4: 20 most frequent 4-gram types in LINDSEI_CZ

4 out of 20 most frequent 4-gram types extracted from LINDSEI_CZ include repeats and/or hesitation items - this may be observed just by looking at Table 4. By sorting through the tokens included in the types of 4-grams and their immediate context in the corpus, however, it is possible to find that 11 of the types include at least one token containing a repeat and/or a hesitation item (e.g. *I I would like to*). In the case of LOCNEC, the easily observable repeats and hesitation items appear also in 4 out of the 20 most frequent types. However,

[8] De Cock uses chi-square test for her values.

there are also altogether 11 types which include at least one token containing a hesitation item and/or a repeat.

Rank	4-gram	Rank	4-gram
1	I don't know I	**11**	or something like that
2	it was it was	**12**	I don't know if
3	and things like that	**13**	it was really good
4	erm I don't know	**14**	a lot of people
5	at the end of	**15**	and it was really
6	the end of the	**16**	but I don't know
7	I think it was	**17**	and it was like
8	I'd like to go	**18**	I don't I don't
9	yeah yeah it was	**19**	I thought it was
10	a bit of a	**20**	and it was just

Table 5: 20 most frequent 4-gram types in LOCNEC

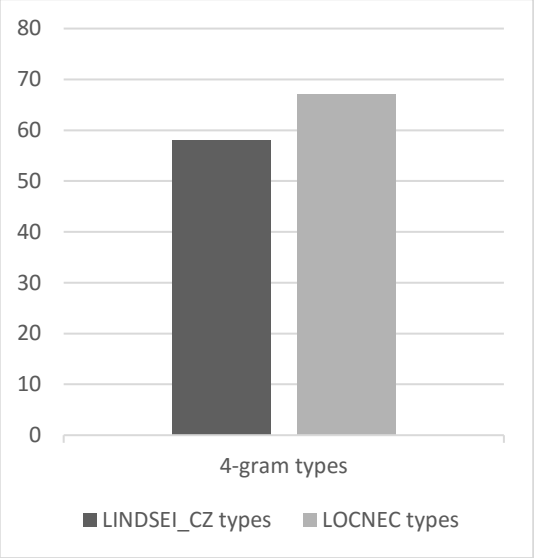

Figure 12: Czech vs native speaker 4-gram types (without repeats and hesitation items)

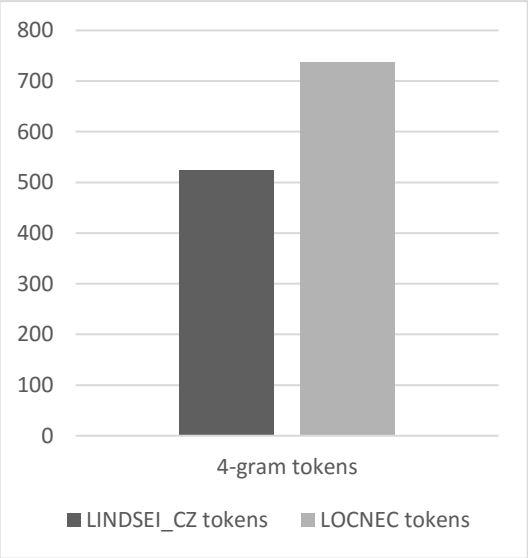

Figure 13: Czech vs native speaker 4-gram tokens (without repeats and hesitation items)

The results in De Cock's analysis where combinations containing repeats and/or hesitation items were left out of the data paint a different picture from the previous results; they show that French speakers use significantly fewer 4-gram types ($p \leq 0.05$) and tokens ($p \leq 0.005$). The next step in the present analysis was then to analyse the data excluding combinations containing repeats and/or hesitation items from LINDSEI_CZ and LOCNEC, to see whether the results show any difference from the comparison above, which included them. The differences are shown to be far less radical than those in De Cock's study, which may be observed in Figures 12 and 13. LINDSEI_CZ types are reduced to 58 per 100,000

word-tokens and LOCNEC types to 67 per 100,000 word-tokens, and the difference between the two groups is again not statistically significant. By removing all tokens containing repeats and/or hesitation items, the relative frequencies are reduced to 524 for LINDSEI_CZ and to 737 for LOCNEC, which shows a statistically significant difference ($p < 0.0001$) between the number of tokens used by the two groups of speakers. The results then show that while native speakers do not use a much greater number of 4-gram types at the set frequency and range, they definitely use them more frequently.

To further illustrate how significant repeats and hesitation items may be in the analysis of recurrent word-combinations, Table 6 shows how many combinations containing repeats and/or hesitation items appear in the overall production of the two groups and also shows the corresponding percentages included in De Cock's study (2004: 232) for the sake of comparison.

Speakers	**Types**	**Tokens**
LINDSEI_CZ	25%	39%
LOCNEC	17%	19%
French speakers (De Cock)	45%	47%
Native speakers (De Cock)	12%	13%

Table 6: Percentages of 4-grams containing repeats and/or hesitation items

One of the important things that should be addressed is the difference in percentages for native speakers in the present study and De Cock's study. This difference is most probably caused by the fact that the present thesis excludes Task 3 of the interviews from the corpora, whereas De Cock includes all three tasks. While the relative frequencies are normalised at the same rate (per 100,000 word-tokens), the results are not exactly comparable precisely due to this exclusion. What may be assumed from the results of the analysis, however, is that Czech speakers are more similar in their frequency of use of recurrent 4-word-combinations to native speakers than French speakers, which might imply a greater language proficiency, or possibly a greater fluency of speech.

4.2 Qualitative analysis of recurrent word-combinations in LINDSEI_CZ and LOCNEC

The following, largely qualitative analysis uses two samples of 15 speakers from each corpus, LINDSEI_CZ and LOCNEC. The chapter is further divided into smaller chapters which correspond to functional categories which are assigned to all types and tokens included in the data extracted from the production of the 30 speakers. The chapter presents examples[9] of the occurrences of word-combination types which are judged to be helpful in illustrating the functional analysis performed on the LINDSEI_CZ sample (henceforth LSS) and the LOCNEC sample (henceforth LCS). The size of the samples is 23,554 word-tokens for LSS and 33,112 word-tokens for LCS. The analysis is also occasionally taken further through additional searches of the whole LINDSEI_CZ and/or LOCNEC corpus (Tasks 1 and 2[10]), providing quantitative data to supply further commentary on certain aspects of the speaker production which might be applicable beyond the two samples of 15 speakers. All quantitative data in this chapter is given in absolute frequencies/number of occurrences, unless stated otherwise.

All tokens of the types which appear in LSS and LCS are sorted into four categories. Tables 7 and 8 show absolute frequencies of functional sequences and the percentages each function makes up in both samples. The data in the tables does not include occurrences of a certain type in cases where the same occurrence also belongs to an overlapping type.[11]

FUNCTION	**N.**	**%**
referential	12	11%
interactional	45	42%
discourse-organising	19	17%
propositional	34	30%
TOTAL	**110**	**100%**

Table 7: Frequency of functional types in LSS

FUNCTION	**N.**	**%**
referential	44	33%
interactional	48	36%
discourse-organising	6	5%
propositional	35	26%
TOTAL	**133**	**100%**

Table 8: Frequency of functional types in LCS

Figure 14 illustrates the differences in frequency of each function in each sample in relative frequencies per 10,000 word-tokens. The differences in relative frequency are not overall

[9] Each example is supplied with a number that corresponds to the respective number of the excerpt in the Appendix and also with a number that corresponds to each speaker in the sample.
[10] The whole of Task 1 and Task 2 is henceforth referred to as T1T2 of LINDSEI_CZ/LOCNEC.
[11] An exception was made for the occurrence where the *I think it was* type overlapped with *it was in the* in the case of one occurrence, because the types themselves serve a different function. (see Chapter 4.2.5) This is why the total in Table 7 does not add up to the number of excerpts in the table Appendix 6.

significant enough to say that one group of speakers generally tends toward certain function more often than the other group. This is most likely due to the small size of the sample. The proportions in absolute frequency do, however, show that in these two samples, the Czech and the native speakers differ in their use of referential word-combinations and discourse-organising word-combinations, whereas they do not much differ in the case of the two other functions. The Czech speakers show a greater need for discourse-organising combinations and the native speakers, on the other hand, for referential combinations. These differences are further discussed in Chapters 4.2.1 and 4.2.3. Both groups use word-combinations that are propositional often, which is to be expected. It is still a fact that both groups use at least 10% more interactional that propositional word-combinations.

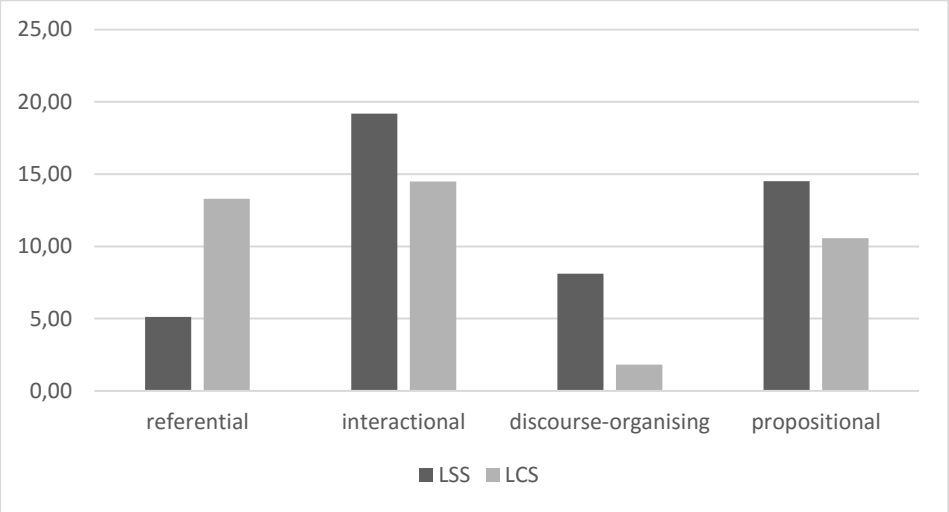

Figure 14: Functions in LSS and LCS (relative frequency per 10,000 word-tokens)

Tables which show the 29 most frequent types in LSS and LCS and their frequencies are available in the Appendix 4. They also show their absolute and relative (per 100,000 word-tokens) frequencies in T1T1 of LOCNEC and LINDSEI_CZ to illustrate their overall standing in a sample of 50 speakers. The most important data in these tables for the following chapters are the absolute frequencies in T1T2 of both corpora. This data is occasionally used in further commentary, usually for comparison, of the use of the analysed types of word-combinations.

TYPE	FUNCTION	N.	TYPE	FUNCTION	N.
I would like to	interactional	13	decided to talk about	discourse-organising	4
	discourse-organising	2	we went to the	propositional	4
something like that	interactional	9	and it was really	referential	3
on the other hand	discourse-organising	8	here in the Czech	propositional	3
in the Czech Republic	propositional	8	I don't know if	interactional	2
it was in the	propositional	6		propositional	1
something like that and	interactional	5	I'm not really sure	interactional	1
	propositional	1		propositional	2
the Czech Republic and	propositional	5	like to talk about	discourse-organising	3
and so on so	interactional	5	a lot of people	referential	2
but on the other	discourse-organising	5	and I was really	referential	2
I think it was	interactional	5	at the same time	discourse-organising	2
so it was quite	referential	5	I have to say	interactional	2
to be able to	propositional	5	so I decided to	propositional	2
and stuff like that	interactional	5	so I went to	propositional	2
but I'm not sure	propositional	3	something like that so	interactional	2
	interactional	1	would like to do	interactional	2

Table 9: Functional distribution of the 29 4-gram types in LSS (number of occurrences)

TYPE	FUNCTION	N.	TYPE	FUNCTION	N.
and things like that	interactional	12	when I went to	propositional	5
I don't know I	interactional	8	a lot of the	referential	4
and it was really	referential	8	a bit of a	referential	2
at the end of	referential	8		interactional	2
it was really good	referential	7	and I was just	propositional	2
I think it was	interactional	5		discourse-organising	2
	propositional	1	I want to do	propositional	4
I thought it was	propositional	6	I was like oh	discourse-organising	4
that kind of thing	interactional	6	it would have been	propositional	4
it was a bit	interactional	2	it's not too bad	referential	2
	referential	3		interactional	2
	propositional	1	know what I mean	interactional	4
things like that and	interactional	5	so I had to	propositional	4
I was going to	propositional	5	you know what i	interactional	4
I went to see	propositional	5	the end of it	referential	4
I'd like to go	interactional	5	the end of the	referential	4
it was a lot	referential	5	a lot of people	referential	3
so it was quite	referential	5			

Table 10: Functional distribution of the 29 4-gram types in LCS (number of occurrences)

Tables 9 and 10 show which categories each of the 29 types of LSS and LCS belong to. The tables already show that certain types are multifunctional. The issue of multifunctionality is

not as straightforward as the tables might imply. This is made more obvious during the analysis of examples in the following chapters and it is specifically addressed in Chapter 4.2.5.

4.2.1 Referential word-combinations

The category of referential word-combinations was inspired by De Cock (2004) and Conrad and Biber (2004). All word-combinations sorted into this category make direct reference to physical/abstract entities or to the textual context by way of modification. Modification may take different forms e.g. quantification (*a lot of the*), intensification (*and it was really*) or specification in time or place (*at the end of the*). Table 11 shows types which appear in LSS and LCS with the referential function.

LSS TYPES	LCS TYPES
a lot of people	and it was really
so it was quite	at the end of
	it was really good
	it was a bit
	it was a lot
	so it was quite
	a lot of the
	a bit of a
	it's not too bad
	the end of it
	the end of the
	a lot of people

Table 11: Referential 4-gram types in LSS and LCS

To begin with, there are markers of quantity in both samples. LCS contains *a lot of* in three 4-gram types extracted – *a lot of the* (ex. 4), *it was a lot* (ex.3) and a *lot of people* (ex. 2) – whereas LSS only contains one – *a lot of people* (ex. 1). There are overall 6 frequent types in T1T2 of LOCNEC (48 occurrences without overlaps) and 3 types in T1T2 of LINDSEI_CZ (23 occurrences without overlaps). There is a single overlap between the types, and that is in LOCNEC; this makes the types unique, if only as 4-gram variations of the *a lot of* sequence. The difference in the frequency of use as such is, however, not significant.

(1) *that **a lot of people** (er) like the Japanese for example they take a lot of photographs* (LS097_ CZ012)

(2) ***a lot of people*** *had said it was really good* (LC132_ EN031)

(3) *for a start* ***it was a lot*** *colder* (LC084_ EN017)

(4) ***a lot of the*** *language has been just translated* (LC099_ EN048)

The placement of the word-combinations *so it was quite, and it was really, and I was really, and it was really, it was really good* and *it's not too bad*, examples of which are directly below, is debatable. The final decision was to consider these borderline referential word-combinations – they might be seen as expressing a personal attitudinal stance to a hearer in interactional contexts, but may only safely be seen as modifiers with no clear interactional function.

(5) *he had: aircondition* ***so it was quite*** *fine for us [because in the rest of India you know we suffered from the heat]*[12] (LS058 _ CZ009)

(6) *it was here in <foreign> Rudolfinum </foreign>* ***so . it was quite*** *an event* (LS059_ CZ014)

(7) *you have to cut it up to produce this booklet .* ***so it was quite*** *difficult .* (LC088_ EN003)

(8) *and* ***it was really good*** *because the different styles of music <\B><B> suited the characters really well <\B>* (LC023_ EN009)

(9) *and* ***I was really*** *quite (eh) . struck by it* (LS099_ CZ011)

(10) *yeah it* ***is it's not too bad*** *at all* (LC118_ EN040)

Example 6, judging from the following context, seems to at the very least present a possibility that the combination may work in interaction as a certain downtoner or hyperbole – there is certainly the hedging potential of the word *quite*. *Really* also has a potential as an amplifier with interactional purpose. None of the available examples, however, present enough contextual evidence to warrant inclusion in the interactional category. De Cock (2004) herself includes these kinds of combinations in her interactional/interpersonal category. She does not actually discuss these combinations or why they should be included in this category.

While there are no more referential word-combination types in LSS, there are more in LCS. There is *at the end of*, which lines up with *the end of it* and partly with *the end of*

[12] Square brackets are used to provide surrounding context which is not included in the list of excerpts in Appendix 6.

the. As in example 11 and 12, the occurrences of *at the end of (the/it)* all mark an endpoint in a period of time (*year/term/film*). As example 14 shows, the preposition *at* may sometimes be replaced by other prepositions, here by *towards* where it also causes a slight shift in meaning).

(11) *and then a project* ***at the end of the*** *year* (LC029_ EN003)
(12) *but* ***at the end of it*** *I mean Brecht doesn't want people to think* (LC030_ EN009)
(13) *when I go back home* ***at the end of*** *term* (LC032_ EN009)
(14) *I just sort of you know <\B><B> work towards* ***the end of the*** *year* (LC130_ EN029)

The use of other prepositions is, however, not frequent enough to show in the quantitative results of the analysis. Furthermore, in LSS, Czech speakers do not use the *at the end of (the)* word-combination at all, and it only appears 4 times in the whole of T1T2 of LINSDEI_CZ, which is significantly less ($p < 0.01$) compared to the 21 occurrences in T1T2 of LOCNEC, marking it as a word-combination preferred, for unknown reasons, by the native speakers of LOCNEC.

Last but not least, there are combinations *it was a bit* and *a bit of a*, which appear with significant frequency in LCS and also T1T2 of LOCNEC. They are analysed together, although they are not fully overlapping, because they both span three distinct categories of function (see Chapters 4.2.2 and 4.2.4 for the other two). Examples 15, 16 and 17 are referential and also examples of quantifying word-combinations.

(15) *sort of newly decorated but (er)* ***it was a bit*** *chilly though* (LC062_ EN040)
(16) *there was something that didn't work about it* ***it was a bit*** *slow* (LC063_ EN042)
(17) *<overlap /> to work with <\B><B>****a bit of a*** *ch= bit of a challenge* (LC100_ EN012)

4.2.2 Interactional word-combinations

The label "interactional" which is used for a great number of different word-combinations included in this chapter is borrowed from De Cock's (2004) study and it was chosen for the simple reason that it seems to satisfyingly express the broad category of word-combinations that perform functions which are geared towards affecting interaction between

speaker and hearer. The category overlaps, but does not fully correspond, to Conrad and Biber's (2004) stance expressions, simply because not all word-combinations included in their category are necessarily interactional; this is shortly discussed further below e.g. concerning the *I don't know if* word-combination. Table 12 shows all word-combinations which appear in LSS and LCS as interactional. Word-combinations included in this category are markers of vagueness, uncertainty, politeness and/or pragmatic downtoners.

LSS TYPES	**LCS TYPES**
I would like to	and things like that
something like that	I don't know I
something like that and	I think it was
and so on so	that kind of thing
I think it was	it was a bit
and stuff like that	things like that and
but I'm not sure	I'd like to go
and it was really	a bit of a
and I was really	it's not too bad
I have to say	

Table 12: Interactional 4-gram types in LSS and LCS

The most frequent word-combination on LSS and also in T1T2 of LINDSEI_CZ it *I would like to*. It does not appear amongst the frequently recurring word-combinations in T1T2 of LOCNEC at all, whereas the 3-gram *I'd like to* does. This difference then does not signal overuse or underuse by either group of speakers, simply a preference for contractions in native speakers' production. The only effect of this is that native speakers may sound less formal.

(18) ***I would like to*** *. finally . (er) go somewhere like to to England or the United States* (LS002_ CZ017)

(19) *so somehow . maybe* ***I would like to*** *try . to get into that* (LS007_ CZ026)

(20) *but maybe later* ***I would like to*** *study . something quite different* (LS014 _ CZ042)

(21) *(erm) I I think* ***I'd like to go*** *when I finish my degree here* (LC082_ EN017)

(22) *and so .* ***I'd like to go*** *over and see <overlap /> what it is* (LC079_ EN003)

All occurrences of *I would like to* and *I'd like to go* have been summarily categorised as interactional. This feature often seems to go hand in hand with expressing something

tentatively, expressing uncertainty about one's own desires. While this might not be the case for example 18 or 22, it is definitely the case for examples 19, 20 and 21, where the tentativeness is reinforced by context (*somehow maybe*, *maybe* and *I I think* respectively).

With *or something like that*, the study moves onto the issue of so called vagueness tags. They are devices of intersubjectivity and "can be seen to play a significant role in informal spoken interactions on an interpersonal level: they signal an assumption of shared experience and social closeness." (De Cock, 2004: 236).

(23) *he's like (eh) .. one thousand years old or* ***something like that . and*** (LS021_ CZ036)

(24) *gave something on it like . to stop the bleeding . like (eh) .. (er) clear sheet* ***or something like that*** *and* (LS019_ CZ017)

(25) *as probably some parents expect that those were just . few words some . family members some animal . colors* ***or*** *<overlap />* ***something like that*** (LS017_ CZ014)

(26) *we didn't have any mobile phones* ***or something like that*** *. and* (LS018_ CZ017)

(27) *I think he was about sixty five seventy* ***something like that and*** (LS047_ CZ026)

Or something like that and *something like that and* (see examples above) often overlap in LSS and mostly then work as vagueness tags – and example of *something like that* which is not preceded by *or* falls into the propositional category (see Chapter 4.2.4). Another example where *or* is missing is example 27 which, while expressing a certain vagueness of expression, maybe even disinterest in being exact, also expresses approximation, working in tandem with the preposition *about*. Example 26 shows inappropriate use of the word-combination. This inappropriate use was also found in the French subcorpus of LINDSEI where "the NNS corpus includes a few instances of the VT *or something (like that)* in inappropriate, i.e. non-assertive, contexts" (De Cock, 2004: 237) There are, however, only two inappropriate occurrences in T1T2 of LINDSEI_CZ. Overall the word-combination is not unique to Czech speakers in T1T2, it is frequently used by native speakers as well. (see Appendix 3)

Two corresponding word-combinations appear in LSS and LCS: *and stuff like that* and *and things like that*.

(28) *just watching the underwater wildlife, hunting for shellfish and .* ***and stuff like that*** (LS069_ CZ012)

(29) *so they've sort of propped it up with bits of metal <\B><B> and put concrete in the centre of it* ***and things like that*** (LC003_ EN013)

(30) *working with the children maybe doing some repairs* ***things like*** *<overlap />****that*** *<\B><B> and playing games with the kids* (LC067_ EN049)

(And) things like that (and) as a 3-gram appears 4 times in T1T2 of LINDSEI_CZ in the range of 4 speakers, whereas in T1T2 LOCNEC it appears 38 times (without overlaps) in the range of at least 16 speakers. *(And) stuff like that* appears 7 times in in corpus T1T2 of LINDSEI_CZ in the range of 7 speakers, whereas in T1T1 of LOCNEC it appears as a vagueness marker 21 times. 19 of those occurrences are, however, produced by the same speaker (LC0019). 47 out of 50 of the native speakers do not use *stuff*, showing an overall significant preference for *things* it this word-combination. As for the Czech speakers in T1T2 of LINDSEI_CZ, there seems to be no significant preference for either.

Another vagueness tag *is that kind of thing* and it appears in LCS (see examples below). It's not used as a vagueness marker by any of the 50 Czech speakers in T1T2.

(31) *people prancing round with tights on and wearing cod pieces and <overlap />* ***that kind of thing*** (LC053_ EN041)

(32) *so what are your favourite films and I say A Clockwork Orange Reservoir Dogs Natural Born Killers* ***that kind of thing*** (LC056_ EN048)

Native speakers, moreover, use the *that sort of thing* tag in T1T2 of LOCNEC (9 times in the range of 6 speakers), which the Czech speakers also do not use. The native speakers then do not have a significant preference for either of these word-combinations, whereas the Czech speakers do not seem to use either of the word-combinations at all. This cannot be broadly applied to the *kind of/sort of* combinations themselves, however, which is obvious from a surface look at the frequencies of *kind of/sort of* as a 2-gram.

	LS FREQ	**LS RANGE**	**LC FREQ**	**LC RANGE**
kind of	117	42	112	24
sort of	64	17	581	40

Table 13: Frequencies of *kind of* and *sort of* in T1T2 of LINDSEI_CZ and LOCNEC

Table 13 shows absolute frequencies from T1T2 of both corpora – it must be kept in mind that these occurrences include the *that kind of thing* and *that sort of thing* combinations. The Table shows that the Czech learners show a *kind of* significantly more frequently than the native speakers but it is significantly more widespread ($p < 0.05$) amongst the native speakers. The native speakers also have a very significant preference for the *sort of* combination. The findings about Czech learners are partly in line with Larsson Aas' study of Swedish learners. "The 2-word combination *kind of* is overused by the Swedish learners, whereas the longer *(that) kind of thing* was found more often in the native speaker data." (Larsson Aas, 2011: 133) While the Czech speakers do not underuse *kind of,* they do use it frequently enough, and so "this may indicate that [*(that) kind of thing*] is not holistically stored" in learner's memory systems, even though *kind of* is "most likely left without internal analysis." (Larsson Aas, 2011: 133)

A word-combination that should be mentioned immediately after *kind of* is *(it was) a bit of a*. All 2 occurrences of the *it was a bit* type which are interactional and work as a vagueness marker also belong to the *a bit of a* type, indicating that the interactional function is actually rather more significantly carried by the latter type, which also nicely shows the parallel with *kind of.*

(33) *it was **a bit of a** cheat really* (LC058_EN009)

(34) *I mean it was **a bit of a** fabricated situation* (LC060_ EN009)

Examples 33 and 34 show the *a bit of a* combination used to mitigate the following classification of situations or actions as *a cheat* and as *fabricated*, thereby reducing the negative implications. The *a bit of a* word-combination is not significantly more frequent in T1T2 of LOCNEC, but it is significantly more ($p < 0.01$) widespread amongst the native speakers - 17 speakers use it as compared to 4 speakers in T1T2 of LINDSEI_CZ.

The last vagueness tag to be mentioned is *and so on so*, which appears in LSS. The word-combination does not appear in T1T2 of the LOCNEC corpus at all. The 3-*gram and so on* does, but only twice. De Cock notes in her study that "*and so on* and *et cetera*, have been found to be mainly used in formal talk" and that it "adds to the impression of detachment and formality [learners] may well give in informal situations." (2004: 236)

(35) *so: it really also impressed me (eh) the the . Renaissance painters **and so on so:** . I also like this o= or also sculpture because* (LS049_ CZ011)

(36) *how should you write it . and after you after you hand it in how should have you written it* ***and so on . so*** *. but (er)* (LS052_ CZ017)

(37) *I could have . (er) lost my hand or . get really severely injured . lose couple of fingers* ***and so on so*** *maybe it was* (LS050_ CZ017)

De Cock mentions a 3-gram *and so on*. There is therefore the matter of the final *so* which follows in the 4-gram *and so on so*. Speaker CZ011 actually uses *so:* 13 times in Task 1 and Task 2 of their interview (see example 35), either at the end of a phrase or on its own - it does not work as a conjunction and it is startlingly similar to a filled pause. *So:* in these cases seems to be present just to fill the silence until the speaker decides how to continue or to fill the silence when the speaker is not sure whether they want to continue at all. It may also be a case of an individual "not knowing how to end sentences", which would just be a habit or a quirk of someone's individual speech production if present in abundance. 23 speakers in the corpus actually use so: at least once as a sort of filled pause. In T1T2 of LOCNEC there are only 4 speakers who use so: in such a way and none of them more than once. While the occurrences of *so* (not lengthened) may work in a similar way, if not the same (as in ex. 36), the data was not searched for them.

Another word-combination that has its place in the interactional category is *it's not too bad*. In spite of its similarity to the word-combination *and it was quite*/*and it was really*, which were summarily placed in the referential category, some occurrences express enough interactional potential to warrant their placement here.

(38) *yeah* ***it's not too bad*** *but i= I mean if you . occasionally* (LC120_ EN049)

(39) *it's it's thirty-five pounds which is <\B><B>* ***it's not too bad*** *really <\B>* (LC121_ EN049)

In these examples the word-combination servers as a hedge, and each occurrence is further reinforced by context. Example 38 shows the combination as a concession used to mitigate the illocutionary force of the following statement and example 39 is further reinforced by *really*.

I think it was is a word-combination which appears in both samples. In LSS, the Czech speakers always place it medially (see ex. 40) and use it to expresses uncertainty about the truthfulness of the following proposition. According to Aijmer, "when *I think* is not placed first it always expresses uncertainty." (2004: 184)

(40) *and it was even amazing that **I think it was** the tour guide in Globe she described* (LS057_ CZ048)

(41) *it was in a a quite a small place as well **I think it was** the <name of a place> Warehouse in London which is . more like a studio* (LC041_ EN009)

(42) *and **I think it was** . supposed to be just a bit in the future* (LC044_ EN027)

This seems to be the case for both groups of speakers, as the medially placed occurrences all fulfil this function, even in LCS (ex. 41). Example 42 is special in that it is the only occurrence in LCS where the combination is placed initially but is potentially used to mitigate the following proposition.

Another word-combination which appears in LSS is *but I'm not sure,* along with *I'm not really sure*. The occurrences of these types are either categorised as propositional or interactional.

(43) *I thought that girls were better at language and boys were . better at walking <overlap /> **but I'm not sure*** (LS075_ CZ036)

(44) *[<B> I think it would be too small for me now after so many years in Prague </B>*

<A> (mhm) (mhm) so you you think that you gonna need Prague yeah for your adult life </A>

*<B>] well **I'm not really sure** maybe one day when I have children and family* (LS092_ 33)

Example 43 shows the word-combination used as a device to reduce the imposition of the speaker's opinion. The combination in example 44, while possibly simply propositional, might also work as a more polite or hedged rejection of an idea presented in the preceding context. Moreover, it can be left out without changing the meaning of the utterance – maybe carries the meaning of (im)probability just as well on its own; this is not true for the occurrences placed in the prepositional category (see Chapter 4.2.4).

Another interactional word-combination is *I don't know if*. The line between the understanding of *I don't know if* as simply referencing insufficient knowledge and between not wanting to commit to the truth of the proposition is fairly thin. The two occurrences which have been deemed interactional are in examples below.

(45) *[everybody asks you about their homework so you check it </B> </B>] and say oh you're so brilliant you'll speak . **I don't know if** there's a group work . you will speak on behalf of the entire group* (LS089_ CZ046)

(46) *there's (eh) Rose Theatre nearby <overlap /> **I don't** </B><B> **know if you** know* (LS090_ CZ048)

Example 45 rather a unit disconnected from the following if-clause. The utterance is paraphrasable as *everybody [will] say, if there's group work, you will speak on behalf of the entire group*. I don't know then works more as a filler – this is addressed further in this chapter when speaking about *I don't know I* (specifically examples 49 and 50). Example 46 shows the combination used as an indirect question, showing concern for shared background knowledge. (Conrad and Biber, 2004: 68) or also simply expressing concern about contact with the hearer.

A seemingly similar type which, however, appears in LCS, is the word-combination *I don't know I.* All occurrences of this type express uncertainty (ex. 47 and 48) or are used as a sort of filler (ex. 49 and 50). Both, but especially the filler occurrences serve to buy the speaker time a they process their thoughts or as they think of what to say next.

(47) *wanted to go back there . and .. **I don't know I** like life there* (LC017_ EN013)

(48) *drawn by the German mentality and .. **I don't know I** just really enjoyed* (LC016_ EN013)

(49) *it's a city but . but **I don't know I** mean I suppose* (LC018_ EN029)

(50) *yeah possibly .. **I don't know I** n= I've not really thought about the future* (LC019_ EN029)

Aijmer (2004) considers this the use of *I don't know* a filler and uncertainty device in her sample from LINDSEI_SW. From quantitative results, Aijmer also claims that *I don't know* is predominantly a feature of learner language as it is used more than by native speakers. Data from T1T2 of LOCNEC and LINDSEI_CZ does not support this claim for Czech speakers, at the very least not in the case of the *I don't know I* word-combination. Accepting repeats and hesitations, in this case because we're considering the function of a filler and uncertainty device, there are 17 occurrences (by 8 speakers) in T1T2 of LINDSEI_CZ and 56 occurrences of (by 30 speakers) in T1T2 of LOCNEC. This is a markedly significant

difference in frequency ($p < .001$) and in range of speakers ($p < .001$). The study does not take the time to analyse every instance of the word-combination *I don't know* to properly investigate this issue (i.e. checking for over-turn-border cases and analysing for function), but below follows a different attempt to further support the claim to a difference between Czech and Swedish learners.

The rough quantitative results show that there are 208 occurrences of this type in T1T2 of LOCNEC and 148 in T1T2 of LINDSEI_CZ. The data in Table 14 shows the absolute frequency counts of 4-gram types which were extracted during the quantitative part of the analysis of this study (with repeats/hesitation items and overlapping occurrences removed). Any of these might possibly work as a device of uncertainty or as a filler (as evidenced by example 45) but do not necessarily have to.

I don't know TYPE	**LS N.**	**LC N.**
I don't know if	9	15
but I don't know	0	9
I don't know it	0	8
so I don't know	0	6
I don't know what	8	7
and I don't know	5	0
I don't know why	5	0
TOTAL	27	45

Table 14: Frequencies of 4-grams containing *I don't know* in LINDSEI_CZ and LOCNEC

Even if all of these types did not work as either device, the difference in frequency of use of *I don't know* as potential occurrences of the devices would not significantly change for either of the groups of speakers. What is clear, then, is that not only do the Czech learners not use *I don't know* more frequently, they also use the one identified uncertainty/filler device (*I don't know I*) significantly less frequently than the native speakers.

A word-combination unique to LSS and T1T2 of LINDSEI_CZ (8 occurrences/6 speakers), meaning it does not appear at all in T1T2 of LOCNEC, is *I have to say*.

(51) *but I'm enjoying the classes **I have to say*** (LS103_ CZ020)

(52) *[<A> how much looking after did the nine year old girl need </A><B>] well she was an only child and quite spo= spoiled **I have to say** so* (LS104_ CZ033)

I have to say functions in examples 51 and 52 as a signal that the previous proposition is surprising or unexpected. *I have to say* might also be considered a bit more formal than usual, for a fairly a casual interview, which may explain the complete lack of usage by the native speakers and might, again, be a sign that learners tend to sound more formal or bookish.

Finally, the word-combinations *you know what I* and *know what I mean* found in LCS are completely overlapping and on a closer look very clearly function as a 5-word combination, although occasionally preceded by *do*. They all show speaker-hearer interaction in that the speaker uses the combination to trigger stronger connection with the hearer, much like with example 46 of *I don't know if*. Indeed, all 3 occurrences, just like in example 53, are followed by a *yeah* or an *unfilled pause* as a response/reaction to this signal from the interviewer. It may be argued that example 53 is propositional in that the speaker is genuinely asking for an answer and would not proceed without it. This could be determined by the intonation of the utterance, but as it is, it keeps its interactional potential.

(53) *and university was really kind of cotton wool arena* ***do you know what I mean*** [<\B>
<A> *(mhm) yeah* <laughs> <\A>] (LC122_ EN034)

4.2.3 Discourse-organising word-combinations

The title of the category of discourse-organising word-combinations is fairly self-explanatory. The word-combinations included here help organise text by expressing simultaneity (*at the same time*), contrast (*on the other hand*), or by introducing a topic (*decided to talk about*) or a quotation (*I was like oh*). All word-combinations included in this category are listed in Table 15.

LSS TYPES	**LCS TYPES**
I would like to	and I was just
on the other hand	I was like oh
but on the other	
decided to talk about	
like to talk about	
at the same time	

Table 15: Discourse-organising 4-gram types in LSS and LCS

The word-combination *on the other hand* (partly overlapping with *but on the other*) serves as a marker of contrast. It is used 22 times in T1T2 of LINDSEI_CZ by 22 speakers,

whereas it is not at all significantly frequent in T1T2 of LOCNEC. In fact, the word-combination only appears 2 times in the whole corpus, each time uttered by a different speaker. It should be mentioned that in LSS, 3 out of 5 occurrences are preceded by *but* (as in examples 32 and 17) and in the whole of T1T2 of LINDSEI_CZ, half of the occurrences are also preceded by *but*. Clause-initially placed combinations seem to be more likely to include the conjunction, but it is not a rule supported by all occurrences.

(54) *which is nice <u>but</u>* ***on the other hand*** *it's also little bit weird* (LS032_ CZ033)

(55) *I've seen the Lord of the Rings* ***on the other hand*** *. like sixty times in my <starts laughing> life* (LS028_ CZ011)

Another word-combination found in LSS is at the same time and it expresses simultaneity, giving preceding and following context the same value. At the same time actually appears in T1T2 of LOCNEC; 7 occurrences by 4 speakers, compared to LINDSEI_CZ'S 13 occurrences by 10 speakers. This difference is not statistically significant.

(56) *it's exciting and* ***at the same time*** *really soothing and calm* (LS101_ CZ012)

(57) *but it's not* ***at the same time*** *it's not very heavy* (LS102_ CZ041)

It should also be noted that Conrad and Biber (2004) categorise this word-combination differently – they place *at the same time* in the referential category (time reference). They do, however, note that combinations such as these might be multifunctional.

Example 62 is an example of the combination *I would like to*, occurrences of which have been sorted into the interactional and propositional category. Examples of the *like to talk about* combination (ex. 59 and 60) include the very same combination, only with the contraction *I'd*.

(58) *I have* ***decided to talk about*** *the . second topic* (LS080_ CZ042)

(59) *pleasure to meet you as well* ***I'd like to talk*** *about two countries* (LS094_ CZ012)

(60) *and* ***I'd like to talk*** *about my experience in Finland* (LS096_ CZ020)

(61) *to begin with . begin with it* ***I would like to*** *<u>say</u> that . (er) it was my sister who showed it to me* (LS010) _CZ036

(62) *if you don't mind* ***I would like to*** *. actually talk about a series Doctor Who* (LS009)

All these word-combinations, including *decided to talk about* (ex.58), serve the same function, more prominent than any possible secondary interactional function that may be present at the same time (e.g. politeness in ex.62). The function itself is, as can be observed from the examples, topic-introduction. The combinations all help to frame the speaker's discourse. Example 61 is the only possibly borderline occurrence sorted here; it does not necessarily serve to introduce the whole topic the speaker will be talking about, but in combination with the verb *say* and the preceding *to begin with*, it forms a long utterance which is used to introduce the topic while also already providing certain information about it (*it was my sister who showed it to me*). It then becomes clear that what all discourse-organising occurrences containing *I would like to* have in common are verbs of speaking (*talk* and *say*).

If all occurrences in T1T2 of the LINDSEI_CZ corpus are combined and added up, they make up 24 occurrences of topic-introducing word-combinations: 4 of *I would like to (talk about)*, 12 of *decided to talk about* and 8 of *like to talk about*. The only topic-introducing word-combination out of the three included above appearing in T1T2 of LOCNEC is *I'd like to talk (about)* – it is used 2 times, each time by a different speaker. Although it must be kept in mind that there might be other word-combinations with this specific function which have not been discovered during the analysis, using the word-combinations acquired and comparing the absolute frequencies of this framing device in T1T2 of both corpora, the difference in use is significant ($p < 0.0001$). These results further support the claim that learners tend to sound more formal - "rather bookish and pedantic" (Channell, 1994: 21) - to native speakers.

Other 4-gram types which work as discourse-organising combinations are *I was like oh* and *and I was just*, extracted from LCS. In the case of *I was like oh*, the function is immediately clear (see examples 63 and 64). *Like* fulfils an organising function similar to reporting verbs, that is that of introducing a kind of quotation. Müller (2005: 164) states that *like* most frequently adopts this function when appearing in the BE+like construction, which is exactly the case in this word-combination. *And I was just* clearly only fulfils this function when followed by *like*.

(63) *my mom was fussing over me and **I was like oh** god you know* (LC110 _EN003)

(64) *she ran out in tears and **I was like oh jeeze** you know* (LC112 _EN034)

It has been said that "like is frequently used for introducing expressions of emotions and for introducing thoughts or potential utterances" (Müller, 2005: 226). The above occurrences are precisely the kind of quotative constructions which introduce past thoughts as well as expressions of emotions. The word-combination may have a secondary interactional function, as according to Müller (2005: 200), the reporting construction gives processing instructions to the hearer, and that is that they should expect a loose fit between the utterance and the thought it represents. *I was like* as a 3-gram and as a quotative word-combination is present in T1T2 of both corpora and therefore is not a feature of the native speaker production only.

4.2.4 Propositional word-combinations

Chapter 4.2.4 shortly presents word-combinations which were judged as purely propositional, as least in certain occurrences. This means they do not seem to express any concern for the interaction with the hearer, do not organise discourse, are not referential and also do not seem to fulfil any other similar function. Table 16 shows all propositional combinations extracted from LSS and LCS.

LSS TYPES	LCS TYPES
in the Czech Republic	I think it was
it was in the	I thought it was
the Czech Republic and	it was a bit
to be able to	I was going to
we went to the	I went to see
here in the Czech	when I went to
I don't know if	and I was just
I'm not really sure	I want to do
so I decided to	it would have been
so I went to	so I had to

Table 16: Propositional 4-gram types in LSS and LCS

There is not much to say for many of the word-combinations. Combinations such as *in the Czech Republic (and), (we) went to (the)* or *so I decided to* are simply what they seem: all are declarative propositions with no additional function. As they are all listed in Table 16, most will not be listed again in this chapter as they were not rigorously analysed. After a

short discussion of one of the purely propositional word-combinations, Chapter 4.2.4.1 discusses multifunctional types; certain occurrences of these types have been shown to belong in one of the other three categories and only a few are propositional.

All occurrences of the *to be able to* word-combination in LSS have been sorted in this propositional category and all of them are of the modal verb of ability and hold no specific interactional, referential or discourse-organisational value. The word-combination is also not unique to LSS or LINDSEI_CZ.

(65) *we started snorkelling I always had this urge . **to be able to** stay there longer* (LS063) _ CZ012

The only occurrence of interest may be example 65, which shows an inappropriate, or rather erroneous use of the word-combination – one cannot have an *urge* to have an *ability*.

4.2.4.1 Multifunctional word-combinations as propositions

This short chapter presents occurrences of types which were previously shown to fulfil one of the other functions for the sake of specific comparison. The first word-combinations were extracted from LSS – *I don't know if* and *I'm not really sure*, which have been shown to have interactional significance in Chapter 4.2.2.

(66) *I'm not really sure whether I should continue **I don't know if** that's the right way for me* (LS088_ CZ026)

(67) *but it involves a lot of travelling . and **I'm not really sure whether** it's compatible with having a family* (LS093_ CZ036)

Examples 66 and 67 show rather personal uncertainty, in the sense of *I wonder whether* In these examples, speakers do not seem to be concerned with the hearer or the truth of the following propositions.

Examples 68 is a lone occurrence in LSS of the combination *something like that* which is just a proposition. It does not supply numerical approximation and it does not function as a vagueness marker of any kind, it merely expresses likeness.

(68) *so we want to try to . make **something like . that** and . try* (LS048_ CZ036)

The easily observable difference from all the other examples of the *or something like that/something like that and* combinations is that the noun phrase is a direct object of the preceding verb and it is obligatory in the syntax; obligatory parts of the syntax cannot fulfil the function of a vagueness tag.

It was mentioned in Chapter 4.2.2 that *I think it was* at times does not express uncertainty when it is placed clause-initially. This is precisely the case of the single propositional occurrence found in LCS. Example 69 shows *I think it was* rather as an expression of belief than uncertainty.

(69) ***I think it was*** *really obvious to them that we were tourists* (LC043_ EN013)

The word-combination *it was a bit* (along with *a bit of a*) has already been sorted into the interactional and referential category. LCS also, however, provides one occurrence of this type that is purely propositional. As is shown in example 70, *a bit* is in this case a fully lexical noun without any additional function.

(70) *because* ***it was a bit*** *we'd done before* (LC059_ EN009)

Finally, *and I was just* has been previously placed into the discourse-organising category in cases where it was followed by *like*. In the rest of the occurrences (2) in LCS, *just* is used with the meaning of *simply* (ex. 71), and while it is possible to claim that *just* may work as an emphasiser, context does not seem to be enough to warrant a different categorisation.

(71) ***and I was just*** *not used to the distances* (LC102_ EN017)

4.2.5 Summary of findings and further commentary

This chapter provides a summary of findings of the analysis and further commentary on certain features of the data. Firstly, it must be mentioned that significant differences between Czech learners and native speakers have been identified. They can be found mainly in two of the four categories – interactional and discourse-organising. Native speakers seem to use in T1T2 a greater variety of vagueness markers, and significantly more frequently. The sequences which are much more frequently that is with statistical significance, used in

T1T2 by the native speakers are *a bit of a*, *that kind of thing*, *that sort of thing* and *things like that*. For Czech speakers it is *and so on so*, which is unique to T1T2 of LINDSEI_CZ. Further findings which do not fully correspond to previous research, such as *I don't know if,* are discussed in Chapter 5 in connection to research of other authors and therefore it need not be mentioned here.

Discourse-organising sequences are a prominent feature in LSS and also in T1T2 of LINDSEI_CZ, based on additional quantitative data presented in the chapter. Whereas native speakers' *I was (just) like* is not unique to their language production, Czech speakers use *on the other hand* and topic-introducing/framing word-combinations with a significantly higher frequency. The abundance of framing devices in their speech makes the Czech speakers sound a fair amount more formal or bookish. This is further reinforced by their frequent use of the polite word-combination *I would like to*, the above mentioned vagueness tag *and so on so*, but also by the word-combination *I have to say*.

The referential category is less telling in terms of any possible generalisations. Native speakers use the *at the end of* /*the end of the*/*the end of it* word-combinations much more frequently than Czech speakers in T1T2. Speakers in LCS used it to reference an endpoint of a film or a term/year. Whether the reason this word-combination is used by them much more frequently is because the group of 50 speakers talked generally more about films or e.g. holidays or exams or whether it is something completely different is not clear from the analysis. Overall the LCS yielded more referential types of word-combinations, but the overall frequency difference is not significant.

The analysis also yielded a number of multifunctional word-combinations. It must be mentioned, however, that multifunctionality as presented in this study is twofold. At times it actually indicates that a specific word-combination may function in a few different ways (e.g. *something like that, I don't know if, it was a bit*), and at times multifunctionality of a corpus type is the result of context, or rather co-text, as is the case of e.g. *and I was just* which, when followed by *like*, becomes quotational. This is caused by overlaps of certain occurrences. Whereas this study summarily treats the types which include these kind of occurrences the same for the sake of clarity, the problem should be at least mentioned. For example in the case of *it was a bit* and *a bit of a* where certain occurrences serve a certain function depending on which parts of these two types overlap. *It was a bit* alone is multifunctional; usually referential, although possibly interactional – when followed by an adjective – *it was a bit boring*, propositional when *bit* is fully lexical – *it was a bit we'd done*. *It was a bit* sometimes as a whole fulfils an interactional function similar to *kind of* by virtue

of being followed by *of a*. The study, nevertheless, does not provide sufficient space to deal with every single instance. The only instance where an overlap was allowed and when an occurrence shared by two types was sorted into different categories was in the case of *it was in the* overlapping with *I think it was.*

- *they built the Globe I think it was in the nineties*

I think it was was categorised as interactional, whereas *it was in the* stayed in the propositional category. *It was in the* as a sequence of four words simply does not carry any other but propositional meaning, and even then it is much less than for example a more fully fledged proposition *in the Czech Republic*. This then means that the propositional category becomes a sort of "other" category in connection to the three other categories, but this is very much in line with the thinking expressed at the beginning of Chapter 4.2.4 where it was stated that certain word-combinations are categorised as propositional because they are not relevant to the speaker-hearer interaction in any way e.g. an occurrence of *I don't know if*, and they are neither referential or discourse-organising. Word-combinations with less clear propositional meaning are categorised on the basis of what they are not rather than what they are, so to say.

5 CONCLUSION

In this chapter reflections are made upon the findings in terms of the research questions posed before the analysis and in the context of some previous studies. Limitations of the method of extraction and analysis are also discussed, in general and as linked to specific conclusions made in the analysis. Some of these were already mentioned previously and are therefore mentioned only briefly.

Firstly, there is the question of whether the research questions posed in Chapter 3 have been answered. The issue of whether Czech speakers use a smaller number of 4-gram types than native speakers and whether they use them less frequently, which is a result that would support the findings of De Cock (2004), was resolved successfully. The results show that while Czech learners do not use a significantly smaller number of 4-gram types at the set frequency and range, they do use them less frequently. The difference between Czech speakers and native speakers is, however, not as significant as between native speakers and

French learners in De Cock's study. Czech speakers are also much closer to native speakers in the frequency of repeats and hesitation items, which is also true for Swedish and Norwegian learners in Larsson Aas' study (2011). The answer to the second research question about the significance of these phenomena in the data would then be affirmative: they have shown to affect the result of the analysis in that the frequency of tokens and number of types is radically lowered when they are removed.

The third research question concerned the kinds of recurrent word-combinations used by Czech speakers and whether they differ from native speakers. This question was also answered in the analysis, and not only in the more qualitative part. The assumptions made based on the quantitative results in Chapter 4.1 about Czech speakers and discourse-organising word-combinations were later supported by findings in the qualitative analysis. No previous study into learner language which would have made any observations about the frequent use of the word-combinations has been found. It would certainly be an interesting feature of learner language to explore in the production of speakers of L1 other than Czech to see whether this is a shared feature amongst learners, like so many things, or whether it might be a unique facet of Czech speakers' speech.

The significant differences between the two groups of speakers then have been found mainly in two categories – the already mentioned discourse-organising category and the interactional category. Smaller variety and lower frequency in the use of vagueness markers by learners is a feature shared by Swedish (Larsson Aas, 2011) and French speakers as well (De Cock, 2004) Another difference which may have been expected to make itself apparent based on previous research would have been the use of *I don't know* (e.g. *I don't know if*, *I don't know I*).. However, the findings show that not only do the Czech learners not use *I don't know* more frequently than native speakers, they also use the one identified uncertainty/filler combination (*I don't know I*) significantly less frequently than the native speakers. This speaks directly against Ajimer's (2004) findings for Swedish speakers and Larsson Aas' (2011) for Swedish and also Norwegian learners. This different results may therefore imply that the overuse of *I don't know* might be a habit of Swedish and Norwegian learners.

Another claim to difference which was made in the study was that Czech learners tend to sound more formal and bookish than native speakers. This was supported by several findings – the learner's use of topic-introducing word-combinations, choice of more formal vagueness markers and other formal-sounding word-combinations (e.g. *I have to say*), and actual underuse of vagueness tags. This observation has been made about learners before

(Channell, 1994; De Cock, 2004; Larsson Aas, 2011). While the support of other studies gives the observations in the present thesis more ground to stand on, it must still be taken into account that these Czech speakers are speaking in a foreign language in a situation where they are being recorded for future research, which may impact their language choices much more strongly than those of the native speakers'. Their relative formality of expression, especially in the case of topic-introducing word-combinations, may very well be a result of different expectations of the situation.

The final research question which was posed was whether all recurrent word-combinations extracted from the data may be considered 4-grams and what makes them so. It cannot be said that the study reached any satisfying, if any at all, conclusions as to the status of word-combinations as 3-, 4- or more-grams. The method of analysis chosen was not focused significantly enough on this issue and the ambiguities encountered during the functional analysis and the process of extraction (method of extraction is addressed below) only raise more questions. It may be said that in this aspect the study failed. It must also be mentioned, however, that it was not an entirely unintentional failure. The real failure was in setting out too ambitious a goal and then when it became clear that any proper consideration of the fourth research question would have required a more extensive study, priority was given almost entirely to the functional investigation of the data.

Most troubles concerning the issue of whether 4-grams are actually 4-grams or not were caused by the method of extraction, so often used in many of the studies mentioned in Chapter 2. At the very least it might be the imperfect software. Nevertheless, the most prominent problem was overlaps – they were dealt with in the qualitative analysis where overlapping occurrences of the same function were removed from the analysed data, but they were left in the overall quantitative data. The only two studies which mention overlaps are Chen and Baker (2010) and Ädel and Erman (2011, who were inspired by the former study; these authors removed all overlaps from the data before starting their analysis. The question is whether this is the way to go. Are all overlapping occurrences 2-grams, 3-grams, 5-grams or even 6-grams? Certainly in the case of *you know what I mean* in the present analysis. But. *I don't know (if)* is slightly different. An example appeared in the analysis where the following clause introduced by *if* did not have anything to do with the function of the 3-word sequence, but there was also another example where it played a role. Refraining from making any assumptions about formulaicity of the 4-gram, it is perfectly natural to wonder whether the 4-gram might not at least pass the criterion of holistic storage and/or retrieval. It is, of course, impossible to decide in the present study, but could it be that the high frequency

of recurrence of precisely this 4-gram could imply that the speaker's production of it does not necessarily go through the process of retrieving the 3-gram *I don't know* and then simply adds on *if* or more words? It would be much easier if the extraction program provided 4- grams in such a way that once a sequence is passed, it is not revisited again and therefore there would be no danger of extracting both *or something like that* and *something like that and* when searching for 4-grams, but that is not the case and it does not seem feasible. And the indiscriminate removal of all overlapping occurrences, or even worse, overlapping types as wholes, does not seem to be a good choice either. The reconciliation of the data with the actual production may be possible through a more rigorous analysis which would be entirely dependent on deeper study of all specific occurrences, the identification of their status as 2, 3, 4 or more-word combinations based on strict criteria. Another consequence of overlapping sequences is pseudo-multifunctionality which was addressed in Chapter 4.2.5.

A fairly obvious limitation of the analysis is the data on which it was performed. The two samples of speakers were small and therefore analysis and arguments presented in 4.2 (qualitative part of the analysis) may be used to speak about the T1T2 of both corpora only with caution; it is a simple fact that only the 29 most frequent types which also appeared with significant frequency for all speakers were analysed. This modifies the data in such a way that some overall more frequent types were left out of the analysis because the chosen 15 speakers did not use them, and that some overall less frequent types which were used by the 15 speakers and which might have been of interest were left out also. Although additional commentary, which occasionally provided quantitative data for T1T2 of both corpora, gives some support to the application of the conclusions to the whole group of speakers, the approach to these must still be carefully critical and, if possible, explored in research which includes a larger sample of speakers.

Although the plan was, at the inception of the idea for this thesis, to also investigate formulaicity and connection between speech rates and use of recurring word-combinations, it quickly became clear that those would be beyond the capacity of the present study. Nevertheless, the examples are numerically identified with the speakers and absolute frequencies of word-combinations used by each speaker are presented in Appendix 5; they might be useful in future research.

Bibliography

Ädel, A., and B. Erman. 2012. "Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach." *English for Specific Purposes 31:* 81–92.

Aijmer, K. 2004. "Pragmatic Markers in Spoken Interlanguage." in *Worlds of words. A tribute to Arne Zettersten. Nordic Journal of English Studies. Special Issue*, 3 (1): 173-190.

Altenberg, B. 1998. "On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations." In *Phraseology: Theory, Analysis, and Applications*, Cowie, A. P. (ed.), 101-122. Oxford: Oxford University Press.

Biber, D. et al. 2000. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brumfit, C. J. 1984. *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy*. Cambridge: Cambridge University Press.

Carter, R., and McCarthy, M. 1995. "Grammar and the spoken language." *Applied Linguistics* 16 (2): 141–158.

Channell, J. 1994. *Vague Language*. Oxford: Oxford University Press.

Chen, Y., and P. Baker. 2010. "Lexical Bundles in L1 and L2 academic writing." *Language Learning & Technology 14 (2)*: 30-49.

Conrad, S., and D. Biber. 2004. "The Frequency and Use of Lexical Bundles in Conversation and Academic Prose." *Lexicographica, Internationales Jahrbuch für Lexicographie* 20: 56–71.

Cowie, A. P. 1994. "Phraseology." In *The Encyclopedia of Language and Linguistics*, Asher, R. E. (ed.), 3168–3171. Oxford: Oxford University Press.

De Cock, S. 1998. "A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English." *International Journal of Corpus Linguistics* 3 (1): 59-80.

De Cock, S. 2004. "Preferred sequences of words in NS and NNS speech." *Belgian Journal of English Language and Literatures* 2: 225–246.

Granger, S. and M. Paquot. 2008. "Disentangling the phraseological web." In *Phraseology: An Interdisciplinary Perspective*, Granger, S. and F. Meunier (eds.). Amsterdam-Philadelphia: John Benjamins Publishing Company: 27-49.

Gráf, T. 2015. "Accuracy and fluency in the speech of an advanced learner of English." Doctoral Dissertation. Charles University in Prague.

Götz, S. 2013. *Fluency in Native and Nonnative English Speech*. Vol. 53. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Larsson Aas, H. 2011. "Recurrent Word-Combinations in Spoken Learner English: A Study of Corpus Data from Swedish and Norwegian Advanced Learners." MA Thesis. University of Oslo.

Leech, G. 2000. "Grammars of Spoken English: New Outcomes of Corpus-Oriented Research." *Language Learning* 50: 675-724.

Müller, S. 2005. *Discourse Markers in Native and Non-Native English Discourse*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Paquot, M. and S. Granger. 2012. "Formulaic Language in Learner Corpora." *Annual Review of Applied Linguistics* 32: 130-149.

Selinker, L. 2014. "Interlanguage 40 years on: Three themes from here." In *Interlanguage: Forty years later*, Han, Z. and E. Tarone (eds.) Amsterdam-Philadelphia: John Benjamins Publishing Company: 221-246.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Skehan, P. 1996. "Second Language Acquisition and Task-Based Instruction." In *Challenge and Change in Language Teaching*, edited by J. Willis and D. Willis. Oxford: Heinemann: ,17–30.

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Stubbs, M. 2007. "An example of frequent English phraseology: Distribution, structures and functions." In *Corpus Linguistics 25 Years on*, Facchinetti, R. (ed.) Amsterdam: Rodopi: 89-105.

Tarone, E. 2014. "Enduring questions from the Interlanguage Hypothesis." In *Interlanguage: Forty years later*, Han, Z. and E. Tarone (eds.) Amsterdam-Philadelphia: John Benjamins Publishing Company: 7-26.

Wood, D. 2010. *Formulaic Language and Second Language Speech Fluency: Background, Evidence and Classroom Applications*. London; New York: Continuum.

Wray, A., and M. R. Perkins. 2000. "The Functions of Formulaic Language: An Integrated Model." *Language & Communication*: 1–28.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Resumé

Předkládaná diplomová práce se zabývá opakovanými slovními spojeními, přesněji je zaměřena na porovnání dvou skupin mluvčích angličtiny, českých a rodilých, na základě 4-gramové analýzy dostupných mluvených korpusů. Práce je celkově zasazena do kontextu předešlých studií zabývajících se jazykem nerodilých mluvčích angličtiny. Ve výsledku analýza poskytuje náhled do rozdílů jazykového projevu českých a rodilých mluvčích a porovnává zjištění s předešlým výzkumem. Práce je rozdělena do pěti kapitol. Po krátké úvodní kapitole následuje kapitola teoretická, která představuje seznámení s mluveným jazykem, žákovským jazykem, jazykovou zdatností, plynulostí a opakovanými slovními spojeními. Poté následuje kapitola metodologická, která popisuje metodologii a materiál využité k výzkumu. Praktická část pak provádí analýzu žákovských korpusů, jejíž výsledky jsou následně shrnuty a diskutovány. Práce je zakončena závěrem, který uvádí výsledky práce do kontextu předešlého výzkumu a popisuje nepřesnosti a nedostatky práce.

Teoretická kapitola nastiňuje přístupy k mluvenému jazyku a krátce představuje aspekty konverzační gramatiky na základě práce G. Leeche (2000). Dále prezentuje oblast výzkumu žákovského jazyka a jazykové zdatnosti se zaměřením na užívání opakovaných slovních spojení a jejich souvislost především s kompetencí plynulosti. Kapitola dále pokračuje přehledem dvou přístupů k opakovaným slovním spojením, frazeologického a frekvenčního, na základě prací Grangerové a Paquotové (2008). Stejně tak poskytuje seznámení s různými pohledy na opakované slovní kombinace (Cowie, 1988, 2001; Mer'čuk a Burger v rámci práce Grangerové a Paquotové 2008). Dále práce nastiňuje předešlý výzkum opakovaných slovních spojení v psaném i mluveném, rodilém i žákovském jazyce a s nimi spojenou terminologii (Altenberg, 1998; De Cock, 1998; De Cock, 2004; Götz, 2013; Larsson Aas, 2014; Wood 2010). V závěru jsou předloženy dvě různé funkční kategorizace (Conrad a Biber, 2005, Wray a Perkins, 2000), kterými byla inspirována funkční analýza v praktické části práce.

Metodická kapitola představuje žákovské korpusy, ze kterých byla čerpána data pro analýzu. Jsou jimi korpus LINDSEI, konkrétně jeho český subkorpus, a porovnatelný korpus LOCNEC, který obsahuje jazykové projevy rodilých mluvčích. Oba korpusy se skládají z padesáti patnáctiminutových rozhovorů s těmito mluvčími. Kvůli předpokládané obsáhlosti studie byly oba korpusy omezeny na dvě ze tří částí každého rozhovoru. Pro účely kvantitativního porovnání mluvčích z hlediska užívání opakovaných slovních spojení byly extrahovány 4-gramy s frekvenčním minimem sedmi dokladů a rozsahem pěti mluvčích za

pomoci programu AntConc 3.4.4w. Extrahovaná data musela být následně upravena, protože některé extrahované 4-gramy přesahovaly hranice výpovědí.

Kvalitativní analýza byla provedena na 4-gramech extrahovaných z menšího vzorku patnácti mluvčích z každého korpusu. Frekvenční a rozsahová hranice extrakce byla nastavena na dvou dokladech a dvou mluvčích. Typy 4-gramů pak byly porovnány se seznamem typů získaných během kvantitativní analýzy na vzorku všech padesáti mluvčích. Pouze typy, které byly obsaženy v obou seznamech, byly zahrnuty do funkční analýzy. Tím bylo lépe zpřístupněno porovnání na úrovni celého korpusu a rozšířena aplikovatelnost závěrů. Metodická kapitola dále zmiňuje problémovost dat z hlediska překrývajících se typů 4-gramů a v závěru předkládá čtyři výzkumné otázky, které jsou řešeny během samotné analýzy.

Cílem první otázky je zjistit, zda výsledky kvantitativní analýzy podporují závěry z předešlých studií opakovaných slovních spojení v řeči nerodilých mluvčích, tedy jestli čeští mluvčí používají méně typů spojení a zda je používají s podstatně nižší frekvencí. Další položená otázka se týká opakování slov (např. *I I don't know* nebo *I was I was*) a váhání (vyplněné pauzy a nedořečená slova) ve vzorku; ovlivňuje jejich vyřazení výsledky analýzy? Třetí výzkumná otázka se týká toho, jaká opakovaná slovní spojení dvě skupiny mluvčích používají a jestli jsou mezi skupinami právě v této oblasti rozdíly. Poslední otázka se týká toho, která spojení se dají opravdu považovat za 4-gram.

Další částí práce je kvantitativní analýza, která byla vystavěna takovým způsobem, aby byla přímo porovnatelná se studií De Cockové (2004). Ukázalo se, že rozdíly mezi českými a rodilými mluvčími jsou statisticky významné jen ve frekvenci užívání slovních spojení, nikoli v počtu jejich typů, a to až po vyjmutí dokladů obsahujících opakování a váhání. V užívání opakovaných slovních spojení a opakování a váhaní jsou pak čeští mluvčí podobnější rodilým mluvčím než mluvčí francouzštiny.

Následující funkční analýza nejprve uvádí data do kvantitativního kontextu, tedy ukazuje poměrné zastoupení čtyř kategorií v projevu menšího vzorku mluvčích. Dále jsou v kapitole detailně analyzovány některé doklady v podkapitolách dělených dle kategorií: referenční, interakční, organizační v diskurzu a propoziční. Kategorie referenční obsahuje typy dokladů, které odkazují na hmotné či abstraktní předměty promluvy či na okolní text; kategorie interakční obsahuje slovní spojení, která mají konkrétní dopad na interakci mezi mluvčím a posluchačem; kategorie organizační v diskurzu obsahuje spojení, která pomáhají strukturovat promluvu, a kategorie propoziční obsahuje spojení, která nemají v interakci žádný jiný význam než propoziční. Konkrétní typy slovních spojení jsou dále v některých

případech prozkoumány na vzorku všech mluvčích z korpusů LINDSEI_CZ a LOCNEC, načež jsou závěry z porovnání využity k popisu nejen menšího vzorku, ale také vzorku všech padesáti mluvčích z každého korpusu. V závěrečné kapitole praktické části práce jsou shrnuty důležité závěry analýzy.

Znatelné rozdíly byly nalezeny v kategoriích interakčních a organizačních v diskurzu. Rodilí mluvčí prokázali, že mají k dispozici rozmanitější zásobu vágních výrazů (*a bit of a, that kind of thing, that sort of thing* a *things like that*) a že je používají častěji než čeští mluvčí. Čeští mluvčí naopak používají více slovních spojení organizujících diskurz, konkrétně výrazů uvádějících promluvu v přímé řeči (např. *decided to talk about*). Čeští mluvčí nadužívají slovní spojení, která zní oproti rodilým mluvčím dosti formálně, a to jak v kategorii interakční, tak v kategorii referenční i organizující v diskurzu. Jde o mnohem formálněji znějící slovní spojení jako je výraz vágnosti *and so on so* nebo další referenční spojení *I have to say* a již zmíněné výrazy uvádějící promluvu v přímé řeči. Objevená multifunkční slovní spojení byla různorodá.

Referenční kategorie slovních spojení byla méně vypovídající. Rodilí mluvčí z neznámého důvodu používají mnohem častěji než čeští mluvčí spojení *at the end of* /*the end of the*/*the end of it* a jejich vzorek také poskytl větší počet typů referenčních spojení, ačkoliv se počtem výskytů v absolutní frekvenci skupiny mluvčích výrazně nelišily. Některá spojení jsou ve skutečnosti multifunkční (např. *I don't know if*), některá pouze spadají do více kategorií kvůli metodickým nedostatkům - již zmíněné překrývající se typy (např. *it was a bit/a bit of a*).

Typy spojení, které byly i propozičně málo vypovídající (*it was in the*), byly řazeny do kategorie propoziční k ostatním spojením jako např. *in the Czech Republic*. Vzhledem k tomu, že do této kategorie byla řazena téměř všechna spojení, která nevykazují žádnou interakční či referenční funkci, ani funkci organizační v diskurzu, je mnoho typů v této kategorii spíš definováno tím, čím nejsou než tím, čím jsou. I přesto je kategorie druhá nejužívanější v obou korpusech.

V poslední kapitole, v závěru, jsou shrnuty nejdůležitější závěry a jsou porovnány s předešlým výzkumem. Čeští mluvčí užívají opakovaná slovní spojení méně často než rodilí mluvčí, ale užívají podobný počet typů, ačkoliv francouzští mluvčí zaostávají za rodilými mluvčími v obou oblastech, a také podstatně více. Čeští mluvčí jsou v tomto tedy podstatně blíž rodilým mluvčím než mluvčí francouzští. V závěru práce také podotýká, že nadužívání spojení organizačních v diskurzu zatím nebylo identifikováno ve výzkumech na skupinách mluvčích jiných jazyků a podněcuje k dalšímu zkoumání s cílem ukázat, zda jsou čeští

mluvčí v tomto výjimeční, nebo jestli k tomuto závěru jednoduše ostatní studie jen nedospěly.

Nečasté užívání vágních výrazů oproti rodilým mluvčím je společné jak českým a francouzským mluvčím (De Cock, 2004), tak i mluvčím švédským (Larsson Aas, 2011). Ačkoliv bylo v předešlých výzkumech dokázáno, že švédští a norští mluvčí nadužívají oproti rodilým mluvčím spojení *I don't know,* v analýze českých mluvčích se toto neprokázalo. Čeští mluvčí používají mnohem více formálně znějících slovních spojení a již zmíněných výrazů uvádějících promluvu v přímé řeči, závěrečná kapitola ovšem podotýká, že by tomu tak mohlo alespoň částečně být kvůli tomu, že nerodilí mluvčí mohou situaci, kdy je jejich projev v cizím jazyce nahráván pro výzkumné účely, vnímat jako mnohem formálnější, než jak by ji vnímali rodilí mluvčí. Tímto závěr shrnuje dostačující odpovědi na první tři výzkumné otázky.

Závěrečná výzkumná otázka týkající se toho, která spojení se dají opravdu pokračovat za 4-gram, nebyla v rámci analýzy dostatečně zohledněna. Práce tedy nedává na tuto otázku jednoznačnou odpověď. Opět zmiňuje překryvy slovních spojení, které tuto problematiku značně ztěžují. Práce podotýká, že je třeba v budoucnu provést důkladnější analýzu, aby se předešlo ukvapeným závěrům. Práce pak předkládá kritiku dat, která byla použita ke kvalitativní analýze. Kvalitativní analýza pracovala jen s 29 typy opakovaných slovních spojení v projevu 15 mluvčích pro každý z korpusů. Ačkoliv se práce pokoušela limitovanou velikost dat zohlednit náhledy do celého korpusu, nedají se závěry bez výhrad aplikovat na korpusy se všemi mluvčími.